| Degree | Year | Name of Author |
|---|---|---|
| PhD | 2005 | GARZA-GARCIA, A. A. |

# Production of disulphide-bonded domains suitable for NMR structure determination: Application to the SRCR domains of the lymphocyte receptor CD5

Alicia Acely Garza-García

University College London

A thesis submitted for the degree of Doctor of Philosophy

January 2005





ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

# Abstract

This work describes the development of a systematic methodology to overcome two of the difficulties commonly encountered when expressing eukaryotic domains in bacterial hosts, namely the failure to obtain folded protein *in vivo* and the low solubility of the expression product. This methodology made possible the production of samples of the first scavenger receptor cysteine rich (SRCR) domain of human CD5, *h*CD5d1, with properties suitable for multidimensional NMR studies.

The SRCR domains of *h*CD5 express in bacteria as insoluble aggregates. The aggregates were purified in order to perform the folding *in vitro*. Optimal conditions for folding were found using a novel systematic screen based on a fractional factorial design. *In vitro* folding yields were assessed using RP-HPLC and non-denaturing PAGE.

The attainment of *h*CD5d1 protein samples of sufficiently high concentration to perform multidimensional NMR was achieved by performing rational apolar-to-polar mutations selected by analysis of a multiple sequence alignment. Eight single residue mutants were engineered and expressed. Four of them had better *in vitro* folding yields than the wild type and a double mutant was constructed by combination of the best behaved single mutations. This double mutant was used to determine the structure of the domain.

NMR experiments at 298 K showed that some regions of *h*CD5d1 undergo conformational exchange on a microsecond to millisecond timescale, hampering the assignment of the resonance signals. Increasing the temperature to 318 K was found to greatly enhance the quality of the NMR spectra and enabled the assignment of more than 95% of the resonances. The solution structure of *h*CD5d1 was determined using standard interproton distance and dihedral angle-restrained molecular dynamics protocols. Forty percent of the residues were found to be in structurally well-defined regions, including all of the regular secondary structure features found for other members of SRCR superfamily. The remaining residues of the polypeptide appear to be distinctly less well ordered.

To those whose presence in my life makes me suspect that I am the most fortunate human in the Universe:

Brígida, Gustavo and Mark

# Acknowledgements

I am indebted to many for the fact that this arduous process is about to reach its conclusion. Not just to those who directly contributed to the work presented here, but also to all those who supported me and gave me reasons to smile and to keep fighting to tame CD5 throughout these years.

I am utterly grateful to the Mexican National Council for Science and Technology (CONACYT) for their financial support during my PhD. I wish to thank them and all the people of México that made this possible. My thanks also to the Graduate School of UCL for providing me with a complementary scholarship for three years.

It is a pleasure to thank my supervisor Prof Paul Driscoll for all his guidance, trust and support. I have learned from him more things than he will be willing to admit.

A mole of thanks to my team mates, Dr Richard Harris and Dr Diego Esposito for assigning the NMR spectra of CD5d1 and doing the structure calculations while I was still trying to tie some loose ends in the wet lab. A bit more thanks to Richard Harris in his position as NMR manager for setting up the NMR experiments and for sharing some of his knowledge on practical NMR with me.

Many thanks to Prof Neil Barclay, Dr Marion Brown and Dr Namir Hassan at Sir William Dunn School of Pathology, Oxford University for providing us with the DNA of the extracellular region of CD5 and CD6.

My gratitude to Prof Penny Handford, Dr Kristy Downing, Dr Jo O'Leary and Vroni Knott, at the Department of Biochemistry of Oxford University for advice on *in vitro* protein folding and inspiring conversation.

Oodles of thanks to Dr Mark Williams for his critical review of my manuscript, his help and relentless discussion throughout my PhD.

I am grateful to Dr Federico del Rio from whom I have learned a plethora of things about NMR spectroscopy, programming and human relationships.

I am very grateful to my lab mates for their friendship, help and support, for the things I have learned from them and for introducing me to the subtleties of English (and Scottish!) idiosyncrasy. Thanks to Andrew S, Beatriz, Brian, David, Diego, Gillian, Mike, Mark J, Matt, Richard, and Roger.

Many thanks to Dr Snezana Djordjevic and my mentor Prof John Ladbury for their useful suggestions and interest in my project.

I would like to thank my parents for their example, support and unconditional love. Thanks to my brother and to all my wonderful Mexican/Dominican/Swiss/Brazilian/ (and now) English family, for always being there for me, waiting for news on my progresses.

On this side of the Pond I have met some wonderful people that have made all the difference in this voluntary exile of mine. Thanks to Abril, Beatriz M., Danielle, Gabriela, Gillian, Manuel, Richard and Virginia, for their enduring support and encouragement. A special thanks to Virginia and the Arechavala-Gomeza household for providing our larder with scrumptious Spanish food that filled not only our stomachs but also our spirits.

Each and every day I am grateful for the friendship of a handful of individuals whose affection remains unaltered despite the enormous distance between us. Thanks to Aubin, Eva, Liliana, Ricardo, Rina and Yuki.

Last but not least, thanks to my husband. He stoically put up with me and my neurosis since those days when I could not make the ligations to work. Thank you Mark, for being who you are and for remaining by my side in this process of becoming who I want to be.

# Table of contents

# List of figures

CHAPTER THREE

IMPROVEMENT OF SOLUBILITY AND *IN VITRO* OXIDATIVE FOLDING YIELD OF THE N-TERMINAL DOMAIN OF HUMAN CD5 BY SITE-DIRECTED MUTAGENESIS 116



CHAPTER FOUR

THE NMR SOLUTION STRUCTURE OF THE N-TERMINAL SRCR DOMAIN OF HUMAN CD5 145

# List of tables

# Chapter One

## CD5, an enigmatic modulator of lymphocyte activation


### Abstract

CD5 is a 67-kDa monomeric transmembrane protein expressed constitutively on T cells and a subset of B cells, although its expression can be upregulated on all B cells by various stimuli. CD5 influences lymphocyte activation by modifying the strength of the intracellular response initiated by antigen engagement. Activation of CD5 can both amplify and reduce the strength of the cellular response. It is not understood how this dual modulation is achieved, but a critical factor appears to be the concentration of CD5 on the cell membrane. The hyper-reactive phenotype of CD5-negative cells suggests that the primary role of CD5 is to dampen the signal of the antigen receptor, therefore raising the threshold for leukocyte response, in order to avoid activation by self-antigens. A few ligands for CD5 have been proposed but all are poorly characterised. CD5 appears to be able to bind C-type lectins and immunoglobulin domains, among others. Soluble CD5 is able to activate B cells, possibly by binding a transmembrane protein involved in T cell-mediated cell activation. CD5 is in close contact with two other important coreceptors: CD2 and CD6; these associations have only recently been reported, and whether they have any biological significance is yet to be elucidated. Finally, the origin, aims and scope of this thesis are summarised.

## Chapter abbreviations and acronyms

**AgR** antigen receptor
**AP-1** activation promoting 1
**APC** antigen presenting cell
**BCR** B cell receptor
**Bcl-2** B cell lymphoma protein 2
**B-CLL** chronic lymphocytic leukaemia B cells
**CD5cyt** cytoplasmic region of CD5
**CD5ext** extracellular region of CD5
**CDR** complementarity determining region
**DAG** 1,2 diacylglycerol
**FRET** fluorescence resonance energy transfer
**GTP** guanine tri-phosphate
**Ig** immunoglobulin
**IL** interleukin
**IP** inositol phosphate
**IS** immunological synapse
**ITAM** immunoreceptor tyrosine-based activation motif
**KO** knock out
**mAb** monoclonal antibody
**MAP** mitogen-activated protein
**MHC** major histocompatibility complex
**pMHC** MHC bound to peptide
**spMHC** MHC bound to self peptide
**NFAT** nuclear factor of activated T cells
**PI3K** phosphoinositide-3 kinase
**PIP2** phosphoinositol 4,5-biphosphate
**PLC** phospholipase C
**PTK** protein tyrosine kinases
**TCR** T cell receptor
**tg** transgenic
**ZAP** ζ-chain associated protein

## Background

Vertebrates possess a group of cells capable of specifically recognising an enormous variety of molecules derived from other 'foreign' organisms. Each one of these cells is stimulated by exposure to a different antigenic agent and their activation initiates a series of defence mechanisms that have evolved in order to eliminate that specific pathogen (Figure I.1). This phenomenon is known as adaptive or acquired immunity. The specific recognition of molecular components of pathogens is achieved via transmembrane glycoproteins generally referred to as antigen-specific receptors (AgRs). Cells that express AgRs are known as lymphocytes, of which there are three classes: B cells, T cells and natural killer cells (natural killer cells do not need to recognise a specific individual antigen to be activated, therefore they are considered as a distinct category). B and T cells are activated by binding to antigen and also receive additional stimuli from receptor-mediated interactions with signalling proteins secreted by other immune cells (cytokines). Once activated, the lymphocytes initiate protein synthesis and undergo mitotic division, a process known as clonal expansion. Clones differentiate into effector cells, whose function is the elimination of the antigen, and memory cells that are able to initiate a rapid response upon a subsequent exposure to the same antigen.

*The B cell receptor for antigen*

The B cell receptor (BCR) is a transmembrane anchored antibody (Ab). Abs are disulphide-bonded glycosylated tetramers of two identical light chains and two identical heavy chains (Figure I.2A). Light chains, are classified in two groups, $\kappa$ and $\lambda$, and are formed by two immunoglobulin domains, each of which consist of ~110 amino acid residues in a disulphide bonded all-$\beta$ fold. The amino acid sequence of the N-terminal immunoglobulin domain of the light chains varies from one B cell clone to another, and it is called $V_L$, the C-terminal domain is relatively constant and called $C_L$. There are five types of heavy chain or isotypes: $\alpha$, $\delta$, $\varepsilon$, $\gamma$ and $\mu$. Each has an N-terminal variable immunoglobulin domain ($V_H$) and three ($\alpha$, $\delta$, $\gamma$) or four ($\varepsilon$ and $\mu$) immunoglobulin domains that do not vary within the same isotype ($C_H$ 1-4). Only Ab isotypes $\delta$ and $\mu$, commonly known as IgD and IgM, are found in the membrane bound form that constitutes the BCR. The cytoplasmic tail of the BCR is not involved in down-stream signalling. Rather, signal transduction through the BCR is accomplished by CD79, a

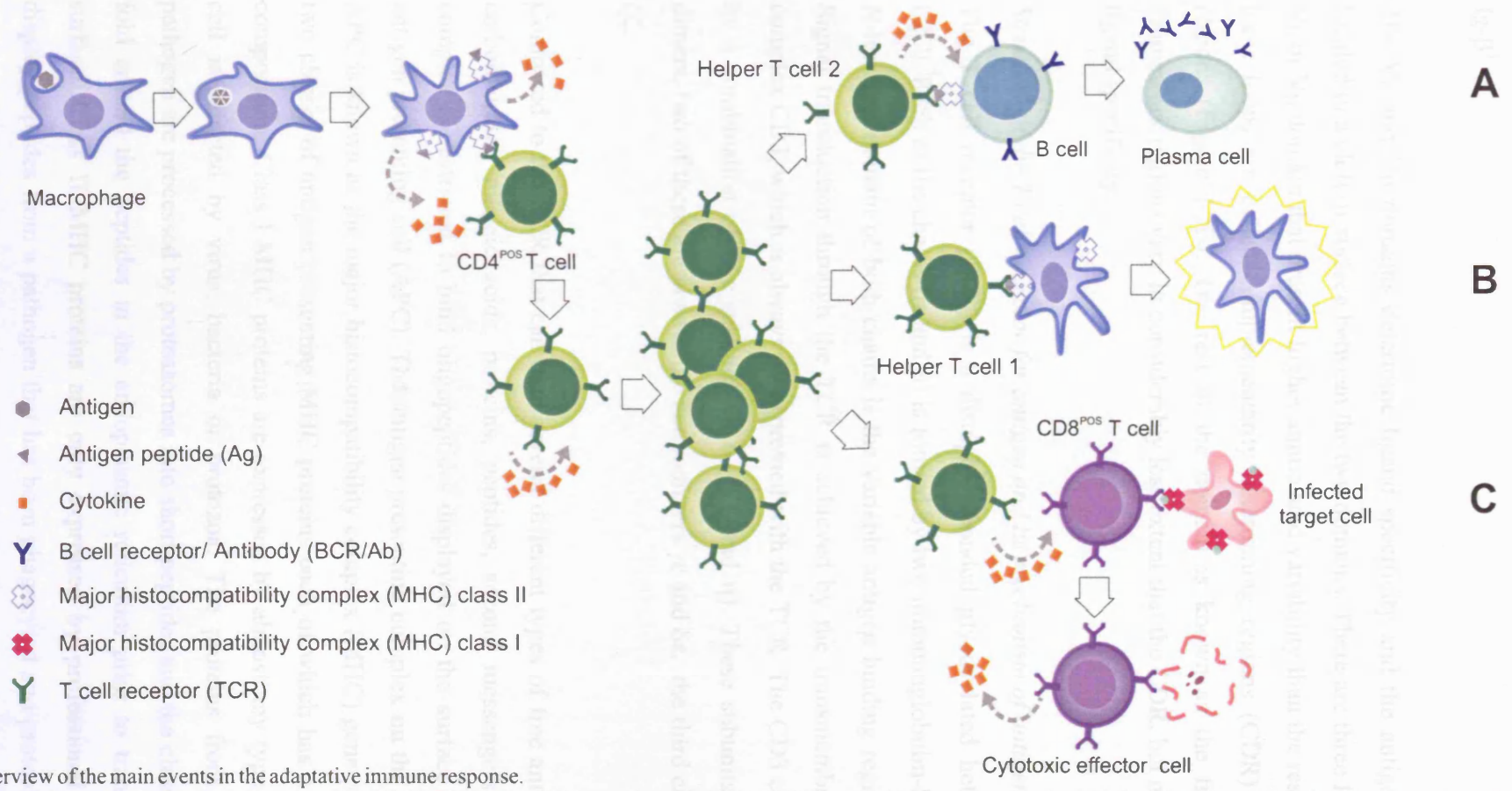


**Figure I.1** Overview of the main events in the adaptative immune response.
Monocytes (macrophages) and other antigen presenting cells (APC) bind and phagocytise pathogens. Pathogen peptide fragments are then displayed on the APC cell surface bound to MHC class II molecules; the MHC-peptide complex is denoted (pMHC-II). APCs secrete cytokines to induce local inflamation leading to increased cellular adhesion and attraction of leukocytes. $CD4^{POS}$ (helper) T cells recognise pMHC-II and become active. $CD4^{POS}$ cells secrete cytokines and undergo clonal expansion. Different types of activated $CD4^{POS}$ cells are involved in subsequent stages of the immune response:
(A) Type 2 helper (Th2) T cells secrete cytokines to promote B cells to divide and differentiate into antibody secreting plasma cells.
(B) Type 1 helper (Th1 or inflammatory) cells signal macrophages to fuse lysosomes and phagocytic vesicles for more efficient pathogen destruction.
(C) Helper T cells may also stimulate $CD8^{POS}$ (cytotoxic) cells. $CD8^{POS}$ cells recognise the MHC-class I molecule, that is expressed on the surface of almost any nucleated cell when it has been infected by intracellular pathogen. Upon recognition of the target antigen, $CD8^{POS}$ T cells become cytotoxic effector cells and induce the infected cell to undergo apoptosis.

disulphide bonded heterodimer whose monomer chains are also referred to as Ig-α and Ig-β[1].

The $V_L$ and $V_H$ domains determine ligand specificity and the antigen binding site is located in a cleft or surface between the two domains. There are three loops in any given $V_L$ or $V_H$ domain that present higher amino acid variability than the rest of the sequence, these loops are called complementarity determining regions (CDR) or hypervariable regions (Figure I.2B). The rest of the domain is known as the framework region. Framework regions vary in considerably less extent that the CDR, but may also influence ligand specificity.

*Structure of the T cell receptor for antigen and its mechanism of antigen recognition*

The T cell receptor (TCR) is a disulphide-bonded glycosylated heterodimer (Figure I.2C). Each of the chains (α and β) is formed by two immunoglobulin-like domains. The N-terminal domain of both chains is the variable antigen binding region (Vα and Vβ). Signal transduction through the TCR is achieved by the transmembrane glycoprotein complex CD3, which is always co-expressed with the TCR. The CD3 complex is formed by a combination of five subunits (γ, δ, ε, ζ and η). These subunits arrange in three dimers, two of these are always the heterodimers γε and δε, the third can either be ζη or ζζ.

Compared to the BCR that can bind several different types of free antigens (*e.g.* lipids, carbohydrates, nucleic acids, proteins, peptides, second messengers), the TCR/CD3 complex is restricted to bind oligopeptides displayed on the surface membrane of an antigen presenting cell (APC). The antigen presenting complex on the membrane of the APC is known as the major histocompatibility complex (MHC) gene product. There are two classes of antigen presenting MHC proteins each of which has a different subunit composition. Class I MHC proteins are expressed by almost any type of cell when the cell is infected by virus, bacteria or protozoa. The proteins from the intracellular pathogen are processed by proteasomes into short peptides and the class I MHC proteins fold around the peptides in the endoplasmic reticulum prior to transport to the cell surface. Class II MHC proteins are only expressed by professional APCs and they display peptides from a pathogen that has been phagocyted and proteolysed. Productive binding of the TCR to the MHC molecule requires other accessory proteins, including

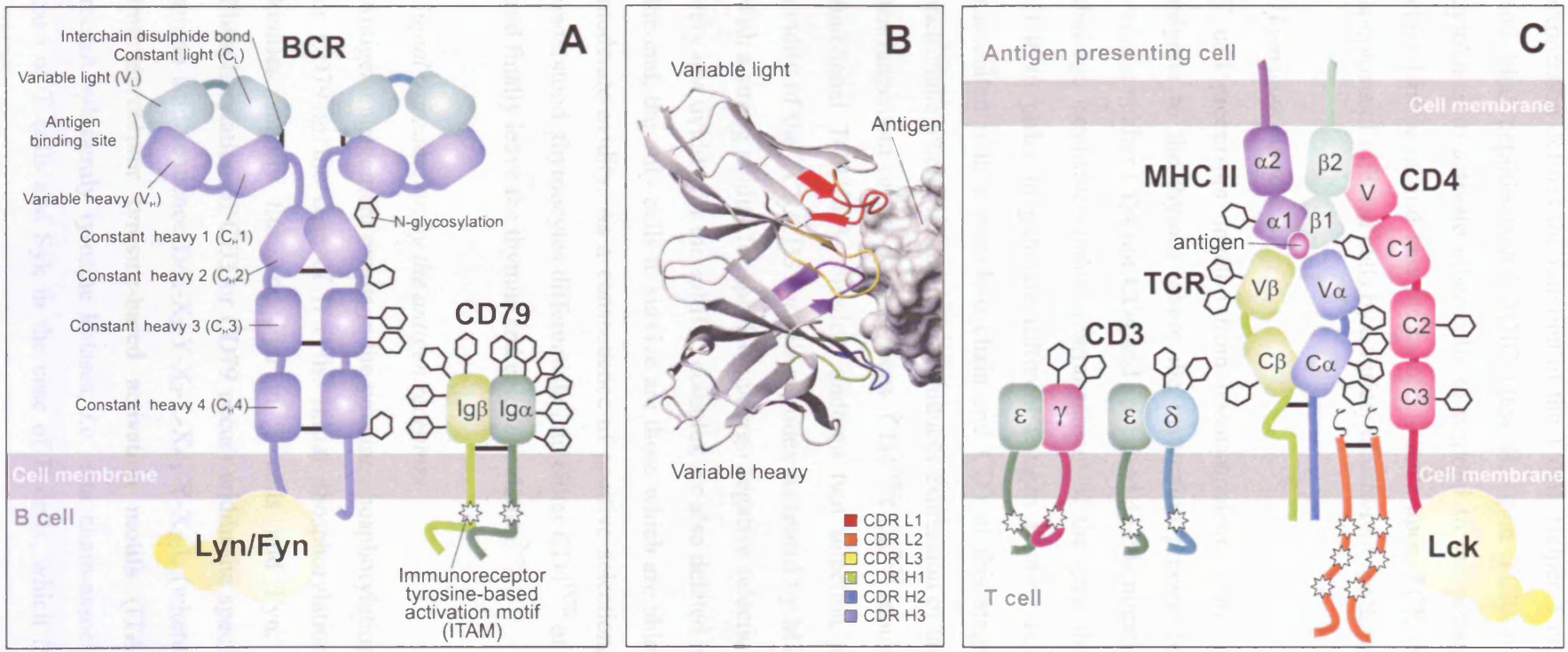




**Figure I.2** Schematic representation of the major immune receptors for antigens.
(A) The B cell receptor is a tetramer of two identical polypeptide heavy chains and two identical light chains. The heavy chains comprise five immunoglobulin (Ig) domains and the light chains two Ig domains. The heavy chains are joined to each other by two disulphide bonds and each heavy chain shares one disulphide with one of the light chains. Signal transduction is accomplished via CD79. The cytoplasmic domains of CD79 contain an amino acid motif with two tyrosines that is phosphorylated by src-family kinases on antigen recognition.
(B) Detail of an X-ray structure of the N-terminal variable Ig domains of both the light and the heavy chains in complex with their corresponding antigen (PDB: 1a2y). The regions which contribute the most to antibody specificity, *i.e.* the complementarity determining regions (CDRs) are highlighted.
(C)The T cell receptor is a disulphide bonded heterodimer. The TCR recognises peptides bound to the major histocompatibility complex (MHC) proteins. Productive binding of the TCR requires the intervention of CD4 (or CD8 in cytotoxic T cells). CD4 binds the C-terminal Ig domain of the β chain of the MHC class II proteins. Signal transduction is accomplished by CD3, which has several cytoplasmic chains having one or three activation motifs that are phosphorylated upon antigen recognition. The src-family tyrosine kinase Lck is associated with the cytoplasmic domain of CD4 (or CD8). The extracellular regions of the BCR, TCR, CD3γ , CD3ε , CD4 and CD79 are formed of Ig domains. The extracellular region of CD3δ is predicted to be highly homologous to an Ig domain, but is not detected as such in databases (PFAM, SMART). MHC class-II is a heterodimer, and each polypeptide is formed by one Ig domain (α2 and β2) and a peptide-binding domain (α1 and β1).

either CD4 or CD8. Each of these binds a different class of MHC proteins and their expression defines the function of the T cell. Helper T cells are CD4-positive ($CD4^{POS}$) and bind peptide-loaded MHC class II; upon TCR engagement these cells secrete cytokines to activate other cells to respond to the pathogen. $CD8^{POS}$ cells bind MHC class I proteins and become cytotoxic cells upon TCR engagement; these cells induce apoptosis of infected cells by a variety of different mechanisms.

*Thymocyte maturation*

T cell precursors develop from haematopoietic stem cells in the bone marrow and migrate to the thymus where the maturation process occurs. T cell progenitor cells express neither CD8 nor CD4 and are called double negative (DN) thymocytes. DN cells undergo extensive random rearrangement of the gene that encodes the β chain of the TCR in order to generate different antigen binding regions. The β-chain, which is associated with a pseudo-α-chain and CD3 at this stage, is transported into the cell membrane where, if functional, it induces expression of both CD4 and CD8, and genetic rearrangement of the α-chain locus. $CD4^{POS}CD8^{POS}$ (double positive cells), now with a functional TCR/CD3 complex, undergo two selection processes that depend on the avidity of the TCR/CD3 for self-peptides presented by MHC molecules (spMHC). Cells with a strong avidity for spMHC undergo negative selection and are deleted. Cells with a very low avidity for the spMHC complex are also deleted in a positive selection stage. At the end, the only cells to survive are those which are able to interact with spMHC with moderate avidity. As a consequence of positive selection, in a process that is not well understood, thymocytes differentiate into either $CD4^{POS}$ or $CD8^{POS}$ (single positive) cells and finally leave the thymus and enter circulation[2-4].

*Signal transduction by the antigen receptors*

Antigen engagement results in the tyrosine phosphorylation of CD3 in the case of T cells or CD79 in the case of B cells. Initial phosphorylation is performed by src-family kinases, mainly Lck and Fyn in T cells and Fyn, Blk and Lyn in B cells[5]. Phosphorylation of CD3 or CD79 occurs within the specific motifs with the consensus amino acid sequence D/E-$X_2$-Y-$X_2$-L-$X_{6-8}$-Y-$X_2$-L (where X is any residue), known as immunoreceptor tyrosine-based activation motifs (ITAM). Phosphorylated ITAMs recruit syk-family tyrosine kinases, *i.e.* zeta chain-associated protein (ZAP-70) in the case of T cells and Syk in the case of B cells, which in turn are activated by trans-

autophosphorylation and by the src-family kinases[6]. Syk-family kinases activate phosphatidylinositol-3 kinase (PI3K) and Tec family tyrosine kinases. Tyrosine kinase activation triggers at least two signal cascades (Figure I.3). One of the cascades involves the activation of phospholipase C (PLC) and the other the activation of small GTPase proteins. PLC hydrolyses phosphatidylinositol bisphosphate ($PIP_2$) into diacylglicerol (DAG) and inositol trisphosphate ($IP_3$). Diffusion of $IP_3$ through the cytoplasm leads to $Ca^{2+}$ release from the endoplasmic reticulum and activation of calmodulin, whist DAG activates protein kinase C (PKC). Small GTPases, such as Ras and Rac, activate the mitogen activated protein (MAP) kinases, including ERK, JNK and p38. The final outcome of the signalling events is the regulation of the cytoskeleton and the activation of various families of transcription factors relevant for lymphocyte proliferation and differentiation. The best known of these transcription factors are: activating protein 1 (AP-1), a heterodimer formed by different combinations of *fos* and *jun* gene products; nuclear factor of activated T cells (NFAT), activated by an increase in the concentration of cytoplasmic $Ca^{2+}$; RelI/NFκB proteins that migrate to the nucleus when their inhibitor is phosphorylated; and the erythroblast transformation specific (ETS)-domain protein Elk-1 activated by phosphorylation by different MAP kinases[5,7-9] (Figure I.3).

## Lymphocyte surface glycoprotein CD5

The precise nature of the lymphocyte response to an activating APC is not controlled by antigen engagement alone, but by the orchestrated response of several accessory transmembrane co-activators and inhibitory molecules. Co-activators stabilise the interaction interface between the APC and the lymphocyte that takes place in a complex structure that has come to be known as the immunological synapse (IS). Within the IS, co-activators amplify and sustain the AgR signal or activate pathways that synergise with the AgR signalling cascade. The degree of activation is regulated by inhibitory molecules that recruit phosphatases, degradation promoting ligases or inhibitory adaptors to the IS[10]. The protein labelled as cluster of differentiation 5 (CD5) is one of many different proteins that participate in the modulation of the AgR response.

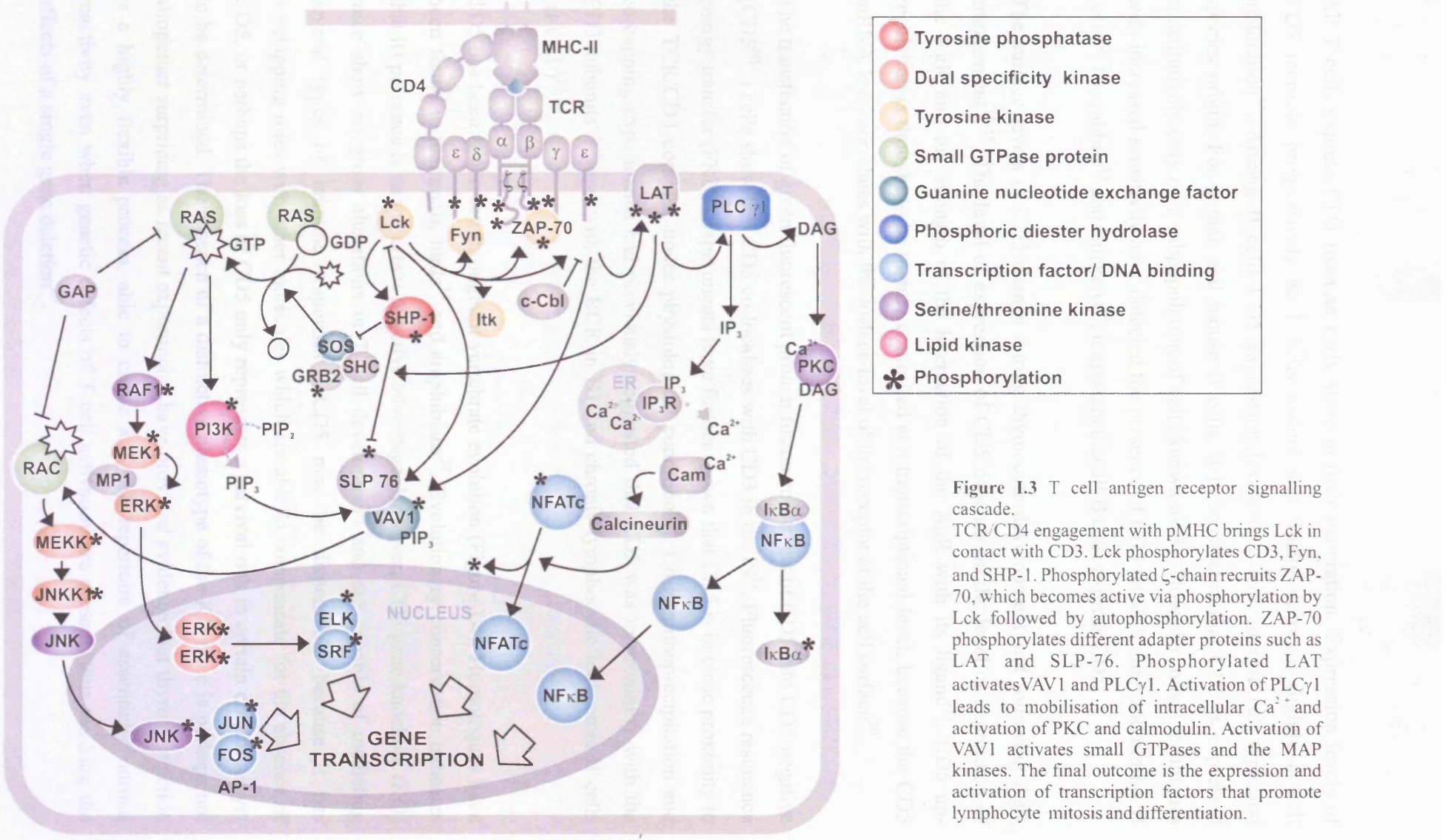


**Figure I.3** T cell antigen receptor signalling cascade.
TCR/CD4 engagement with pMHC brings Lck in contact with CD3. Lck phosphorylates CD3, Fyn, and SHP-1. Phosphorylated ζ-chain recruits ZAP-70, which becomes active via phosphorylation by Lck followed by autophosphorylation. ZAP-70 phosphorylates different adapter proteins such as LAT and SLP-76. Phosphorylated LAT activates VAV1 and PLCγ1. Activation of PLCγ1 leads to mobilisation of intracellular $Ca^{2+}$ and activation of PKC and calmodulin. Activation of VAV1 activates small GTPases and the MAP kinases. The final outcome is the expression and activation of transcription factors that promote lymphocyte mitosis and differentiation.

All T cells express CD5 from an early stage in their maturation. Expression levels of CD5 increase progressively as T cells mature and finally decrease slightly at full maturation[11]. Among B cells CD5 expression levels vary with precise cell type and species origin. For human and mouse B cells, it is considered that CD5 is expressed constitutively only on a subpopulation of cells known as B1. However, novel techniques with increased sensitivity have detected the presence of CD5 on all human and murine B cells[12]. For rabbits[13] and chickens[14], it appears that all B cells express CD5.

The surface levels of CD5 change during thymocyte development and following AgR engagement[15-18]. The level of expression of CD5 depends on the level of expression of the AgR and the strength of the interaction of the AgR with its ligand[19]. CD5 up-regulation is thought to be mainly controlled at a transcriptional level, because the CD5 mRNA level correlates with the surface level of the receptor at the cell surface[20].

The transfection of green fluorescent protein fusion constructs of CD5 into CD5-negative ($CD5^{NEG}$) cells show that CD5 co-localises with CD3 in the IS[21]. Fluorescence resonance energy transfer (FRET) experiments have further shown that CD5 is in close proximity to the TCR/CD3 complex under physiological conditions[22]. Co-immunoprecipitation and co-capping experiments had previously suggested that CD5 was in proximity with the CD3 subunits[23,24] and with the BCR in B1 and chronic lymphocytic leukaemia B cells (B-CLL)[25,26].

CD5 has been conserved throughout vertebrate evolution (Figure I.4). Homologues have been found in mammals, birds[14] and amphibians[27]. Evolutionary conservation indicates that its presence is an important selective advantage. However, CD5 gene knockout (KO) mice show no gross alterations in T cell development and are capable of mediating several types of immune responses[28]. CD5 may be dispensable because it has overlapping roles with other molecules which are able to compensate for the absence of CD5, or perhaps the loss of CD5 only represents a survival risk in certain conditions yet to be determined. The absence of a distinctive phenotype of the KO mice is perhaps not altogether surprising as recent experiments have provided evidence that thymic selection is a highly flexible process, able to create a T cell repertoire of essentially normal reactivity even when genetic defects of T cell activation are present, thus masking the effects of a single gene deletion[29].

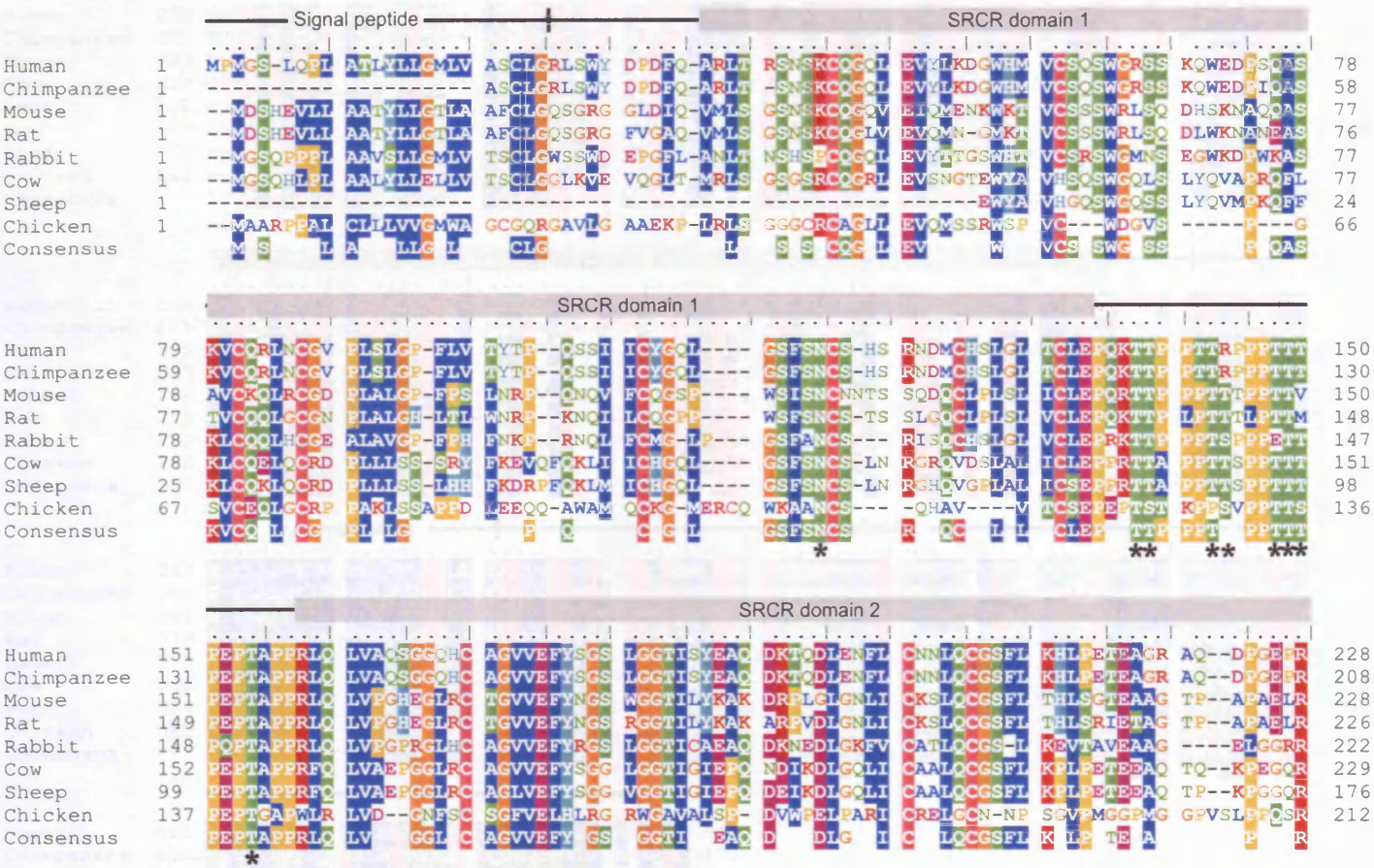


**Figure I.4** Amino acid sequence alignment of human lymphocyte glycoprotein CD5 and some orthologues (continued on the next page).
(*) Predicted N- or O-linked glycosylation site; (▲) Putative phosphorylation site. Red, positive; purple, negative; blue, hydrophobic; green, aliphatic & polar; cyan, aromatic & polar; pink, cysteine; yellow, proline; orange, glycine.


SRCR domain 2
SRCR domain 3
SRCR domain 3
Transmembrane
Cytoplasmic
Cytoplasmic
Human
Chimpanzee
Mouse
Rat
Rabbit
Cow
Sheep
Chicken
Consensus

## CD5 protein structure

CD5 is a type I monomeric transmembrane glycoprotein of 67 kDa formed by a single polypeptide chain of 492 residues (in human). The 348-residue extracellular region contains three ~110-residue domains (CD5d1, CD5d2 and CD5d3) that belong to the scavenger receptor cysteine-rich superfamily (SRCR). The 92-residue cytoplasmic domain has a highly conserved sequence but no discernible 3D fold type. The cytoplasmic domain is predicted to have little regular secondary structure, except for two regions that could form β-strands and an α-helix[30].

CD5d1 and CD5d2 both possess potential N-linked glycosylation sites and are separated by a 25-residue proline and threonine-rich region containing 9 potential O-linked glycosylation sites[31]. O-glycosylation of peptides produces extended structures that are rigid, and it is thought that the CD5d1-CD5d2 spacer region acts as a stalk that helps project CD5d1 away from the cell membrane[32,33]. When expressed as a recombinant protein in Chinese hamster ovary cells, the N-glycan attached to CD5d1 is almost exclusively a fucosylated biantennary glycan (A2G2F) with zero, one or two covalently bound molecules of sialic acid[33] (Figure I.5). It is known that most anti-CD5 antibodies bind to CD5d1, presumably because it is the domain with the highest sequence divergence among species[34-36] (Figure I.4)

## CD5 regulation in lymphocyte activation

The nature of the modulation of AgR signalling by CD5 has been a matter of controversy for many years. At this time there is a fragile consensus that it is primarily a negative modulator, but a good deal of the available data indicate that it can also act as an activation enhancer of leukocyte function. Research groups active in the field appear to be coming to terms with the concept that CD5 is able to both augment and inhibit the responses of the AgR[37] depending on the type of B or T cell, the cellular differentiation state and the nature of the induced stimulus[38,39].

Experimental methodology can alter the cellular response to CD5 stimulation. CD5 is generally activated by addition of anti-CD5 mAbs. The precise effect observed appears to depend on the specific antibody used[40,41] and the method of stimulation[42], *e.g.* cross-linking of anti-CD5 mAbs alone, cross-linking of anti-CD5 mAbs with anti-CD3 mAbs, use of mAbs bound to a solid support. There is also a possibility that some anti-CD5 mAbs might be able to block the interaction of CD5 with its ligand or sequester CD5 away from the antigen receptor complex (as has been done deliberately on some occasions[43,44]), leading to an erroneous interpretation of results[39].

Recent studies have revealed that the level of CD5 on the cell surface has an influence upon the strength of the TCR-mediated response. This observation has turned out to be fundamental to the comprehension of the variable effects of CD5. Thus, it is reported that the concentration of CD5 on the cell surface is inversely correlated with the cell's responsiveness to TCR engagement[18,45]. In turn, CD5 expression level affects the T cell threshold to activation, since low levels of CD5 expression appears to render T cells hyper-responsive[39].

Cell surface levels of CD5 are affected by the avidity of the binding of the TCR to the complex of the corresponding antigenic peptide with the MHC protein (pMHC). TCRs that bind their pMHC with high avidity trigger a strong activating response that increases the expression levels of CD5. Hence, the outcome of direct CD5 stimulation with anti-CD5 mAbs in any particular experimental system would appear to depend on, among other factors, the degree of activation of the cells. The concentration of CD5 on the cell surface is also actively maintained by APCs presenting self-peptides (spMHC)[45-48]. These self-antigens play an important role in the long-term survival and homeostasis of peripheral B and T cells, they have been described as 'ticklers' that maintain a healthy lymphocyte population[49]. Both the frequency of contact and the binding avidity for the spMHCs are reported to influence the expression levels of CD5 in T cells. High avidity and/or a high number of contacts sustain high CD5 expression levels that, in turn, decrease the responsiveness of T cells after pMHC engagement[45].

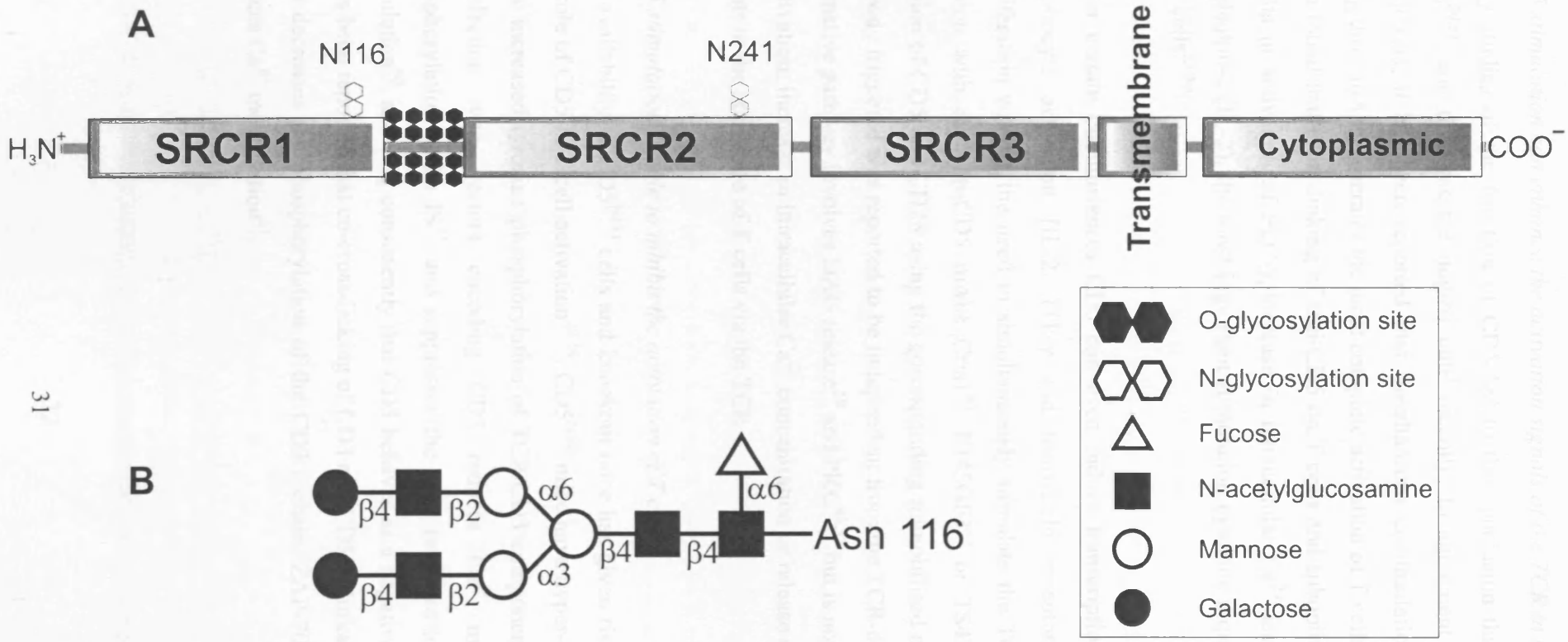




**Figure I.5**

(A) Schematic representation of the general features of the CD5 glycoprotein.

(B) The structure of A2G2F, the predominant glycoform found attached to Asn 116 in CD5d1 expressed in CHO cells. The identity of the other CD5 glycans has not been determined. The greek letter (α or β ) refers to the configuration of the anomeric hydroxyl group in the first monosacharide and the number refers to the position of the carbon attached to the hydroxyl group in the second monosaccharide. All glycosidic linkages represented here originate from the C1 hydroxyl group, *e.g.* an α(1-6) glycosydic linkage is represented as α6.

*CD5 stimulation can enhance the activation signals of the TCR in mature T cells*

Early studies of the function of CD5 led to the conclusion that CD5 co-activates T cells[50,51], and this was the dogma until recently. In agreement with this co-activator hypothesis, it has been reported that simultaneous costimulation of CD3/CD5/CD28 using three mAbs generates the most complete activation of T cells[52]. Additionally, it has been found that cross-linking of anti-CD5 on T cells and suboptimal doses of anti-CD3 results in activation of PLC-γ, increase in intracellular $Ca^{2+}$ levels and expression of interleukin-2 (IL-2), the most important stimulation cytokine required for full activation of T cells[53-56].

Under certain circumstances CD5 can even induce transcription of genes related to lymphocyte activation (IL-2, ITF-γ and transferrin receptor) and cause T cell proliferation without the need to simultaneously stimulate the TCR. Examples include ligation with the anti-CD5 mAbs Cris1[40], F145GF3[40] or TS43[41], and simultaneous ligation of CD5 and CD28 using the corresponding immobilised mAbs[57]. The activation pathway triggered was reported to be independent from the TCR-dependent cascade. The alternative pathway involves DAG release[58] and PKC[40,57] but is not associated with PLC-γ activation, increase in intracellular $Ca^{2+}$ concentration or release of IPs[40], which are key events in the activation of T cells via the TCR.

*CD5 stimulation is able to inhibit the activation of T cells*

The availability of $CD5^{NEG}$ cells and knockout mice has given rise to a reassessment of the role of CD5 in T cell activation[28,59]. $CD5^{NEG}$ mice have hyper-responsive T cells and show increased tyrosine phosphorylation of TCR/CD3 components[59]. In $CD5^{NEG}$ cells, transfection with vectors encoding CD5 reduces $Ca^{2+}$ mobilisation and Tyr phosphorylation at the IS[21], and suppresses the TCR response to antigen or anti-CD3 stimulation[60], showing consistently that CD5 behaves as a negative regulator. Moreover, it has been reported that co-cross-linking of CD3 and CD5 in Jurkat cells (a human T cell line) decreases the phosphorylation of the CD3 ζ chain, ZAP-70, Syk and PLC-γ and reduces $Ca^{2+}$ mobilisation[61].

Dendritic cells (DC) are APCs which are able to suppress peripheral T cell activity against self-antigens by several distinct mechanisms[62]. Recently, it was reported that one of these mechanisms involves inducing T cells to up-regulate CD5 expression and that this process renders T cells anergic *in vivo*, *i.e.* unable to respond to antigen. Despite their unresponsive state *in vivo*, these DC-exposed $CD5^{HIGH}$ T cells could be activated *in vitro*. T cells that had not been exposed to DC ($CD5^{LOW}$) exhibited enhanced activation upon co-cross-linking of CD3, CD28 and CD5 compared to co-cross-linking of only CD3 and CD28, supporting the idea that CD5 acts as a co-activator. However DC-exposed T cells ($CD5^{HIGH}$) showed the opposite behaviour, suggesting that high levels of CD5 down-regulate T cell activation through a pathway independent of the TCR signalling cascade[63].

*CD5 increases the threshold for TCR stimulation during T cell maturation*

The expression of CD5 starts in very early stages of thymocyte maturation and is dynamically up-regulated during thymocyte development[16,17]. Thymocytes are also more responsive to CD5 stimulation than mature T cells[11]. Most studies support the hypothesis that CD5 is involved in raising the threshold for negative selection in thymocytes. Thymocytes of $CD5^{NEG}$ mice show increased proliferation, increased tyrosine phosphorylation of PLC-γ and Vav and increased $Ca^{2+}$ influx[59]. Furthermore, the level of negative selection following stimulation with anti-CD3 mAbs is much higher in $CD5^{NEG}$ mice than in wild type mice[39,59]. All these results suggest that the presence of CD5 dampens the activation responses in thymocytes.

Experiments using transgenic mice that have a particular TCR (tg-TCR) expressed in large numbers of T cells have been very informative as to the effect of CD5 on the outcome of thymic maturation. Thymocytes bearing a tg-TCR that is known to normally undergo severe negative selection can be rescued by activation of CD5 with anti-CD5 mAbs[39]. Conversely, $CD5^{NEG}$ mice show increased negative selection of thymocytes expressing the tg-TCR DO11.10, which in normal mice is expressed in virtually all $CD4^{POS}$ T cells[60]. Both of these results show that very reactive TCRs avoid apoptosis in the thymus if CD5 is active, but otherwise would experience an inadmissible level of activation and would be deleted. Therefore, it seems that CD5 activity reduces the number of thymocytes undergoing apoptosis because of their extremely high levels of

activation (negative selection), but also increases the number of thymocytes that undergo apoptosis because of their unacceptable low levels of activation (positive selection).

A chimeric construct consisting of the extracellular domain of the human receptor Tac, that has no ligand in mice, and the cytoplasmic region of CD5 has been reported to be able to inhibit TCR signalling in $CD5^{NEG}$ thymocytes. This observation suggests that the function of CD5 in the thymus does not require engagement of any ligand[64]. The authors of this study proposed that CD5 acts constitutively as a negative regulator that can switch to being a co-stimulator upon ligand binding in mature T cells. However, these results may simply indicate that there are other receptors in the thymocyte surface able to counteract the effect of CD5's inability to bind its ligand.

Despite the foregoing evidence, there is a degree of controversy concerning the negative regulator role of CD5 in thymocytes. A few reports have suggested that CD5 together with other receptors such as CD28, the main T cell co-stimulator, are responsible for TCR activation in thymocytes and, in this manner, CD5 enhances, rather than reduces, negative selection[65,66]. Two explanations for this apparent inconsistency have been suggested. It could either be that the effect of CD5 is different at different stages of thymocyte maturation, or that the particular mAbs used in order to stimulate CD5 in these experiments are actually having unanticipated effects and acting in the opposite sense to that supposed[39].

*CD5 inhibits the signalling events activated by BCR engagement and promotes the production of a cell survival cytokine*

B cells from humans and mice are classified according to the presence of CD5 on their surface: B1 cells are $CD5^{POS}$ and B2 cells are $CD5^{NEG}$. B1 cells are able to produce Abs (mainly IgM and IgA) in the absence of any given antigen stimulation[67]. These, so-called natural, Abs have low affinity and broad specificity, and are able to recognise bacterial antigens, DNA, multicellular parasites and autoantigens[38,68-70]. The role of B1 cells is not well characterised, but they are known to be involved in the defence against bacteria and parasites[71-73], and in modulating the Ab repertoire by establishing the idiotypic network[74] (an interconnecting series of self-Abs able to recognise the variable regions of the organism's own BCRs). B1 cells are also associated with autoimmune diseases and malignant transformations. Thus, B1 cell-count increases in some autoimmune

conditions such as rheumatoid arthritis, Sjögren's syndrome and lupus[70]. Several B cell malignancies, *e.g.* chronic lymphocytic leukaemia and mantle-cell lymphoma, are characterised by cells that express CD5 at high levels[75,76].

The behaviour of B1 cells is in agreement with a role for CD5 as an inhibitor of the signalling events triggered by the BCR. For example, $CD5^{NEG}$ cells become active and begin to differentiate upon BCR engagement, in contrast the response of $CD5^{POS}$ cells to BCR engagement is poor: cells do not proliferate and instead show enhanced apoptosis[77]. Moreover, $CD5^{POS}$ cells have reduced PLC activation, delayed or absent triggering of MAP kinase cascades and are unable to activate NF-κB transcription factors[78,79]. B1 cells of CD5-deficient mice are resistant to apoptosis after BCR engagement and undergo proliferation[43]. Additional findings consistent with a role for CD5 in negative regulation of the BCR have been obtained by expression of a tg-BCR that renders B cells functionally impaired. In contrast to the tg-BCR/$CD5^{POS}$ anergic phenotype, tg-BCR/$CD5^{NEG}$ cells are hyper-responsive *in vitro*[80], suggesting that CD5 plays a key role in the impaired activity of the tg-BCR cells.

Down-regulation of the BCR response is not the sole function of CD5 in B cells. B1 cells are long-lived and self-renewing because they are able to produce interleukin-10 (IL-10), a potent immunosuppressive cell-survival cytokine[81]. $CD5^{NEG}$ cells transfected with a vector containing CD5 produce more IL-10 than the control mock transfected cells. CD5 was found to activate the promoter of the IL-10 gene without requiring BCR stimulation[82].

New evidence suggests that B2 cells do express CD5 in very low quantities[12]. Moreover, the expression of CD5 in B2 cells can be up-regulated using different stimuli such as anti-IgM antibodies, T cell-independent antigens such as bacterial polysaccharides, expression of retroviral GTPases, and BCR cross-linking[83-85]. These observations have led to the hypothesis that B1 cells are simply an anergic state of B cells caused by chronic stimulation, rather than a separate cell lineage[86]. In mice unable to express the SHP-1 phosphatase (the so-called 'motheaten' mutation) not only are most of B cells $CD5^{POS}$ but macrophages are also found to be $CD5^{POS}$ [87], leading to support another ongoing hypothesis that B1 cells and macrophages develop from a common progenitor cell[88].

To summarise, it is clear that alteration of the activation state and the cell surface concentration of CD5 modifies the strength and the nature of the intracellular response initiated by the antigen receptor of T and B cells. Unfortunately, the precise mechanism of this modification is still far from transparent. The reports from different laboratories, using distinct methods to manipulate CD5, are not obviously consistent and seem to be susceptible to a variety of interpretations.

## CD5 signal transduction

The 92-residue cytoplasmic domain of CD5 (CD5cyt) does not possess intrinsic catalytic activity, but it contains 11 serine residues, 4 threonines and 3 tyrosines, and of these, 10 are predicted to be phosphorylation sites[89]: Ser404, Thr412, Ser415, Ser423, Tyr429, Ser439, Ser458, Ser459, Ser461 and Tyr463* (Figure I.4). Experiments have shown that CD5 recruits both protein kinases and phosphatases and undergoes phosphorylation on serine[90], threonine[91] and tyrosine[23,92,93] residues either constitutively or following stimulation of the AgR and/or stimulation with anti-CD5 Abs[92]. Although current knowledge of the signalling events triggered by CD5 stimulation is very fragmented, it appears that CD5 is able to alter the molecular signals generated as a result of AgR engagement, and it is also involved in AgR-independent signalling pathways.

*Casein kinase 2 phosphorylates S459 and S461*

Casein kinase 2 (CK2) has been found by co-immunoprecipitation and yeast two-hybrid assays to interact with the cytoplasmic domain of CD5. CK2 phosphorylates S459 and S461 upon stimulation with anti-CD5 Abs[94,95]. CK2 is probably the most pleiotropic individual protein kinase[96] and the final consequence of this activation is not yet known. However, there has been some speculation that this pathway could lead to the activation of the AP2 adapter complex that mediates receptor-dependent endocytosis[97].

* Note: Authors that have published on signal transduction involving CD5 do not include the signal peptide (Met1 to Gly24) in their residue numbering. I have followed this convention only in this Chapter of the thesis.

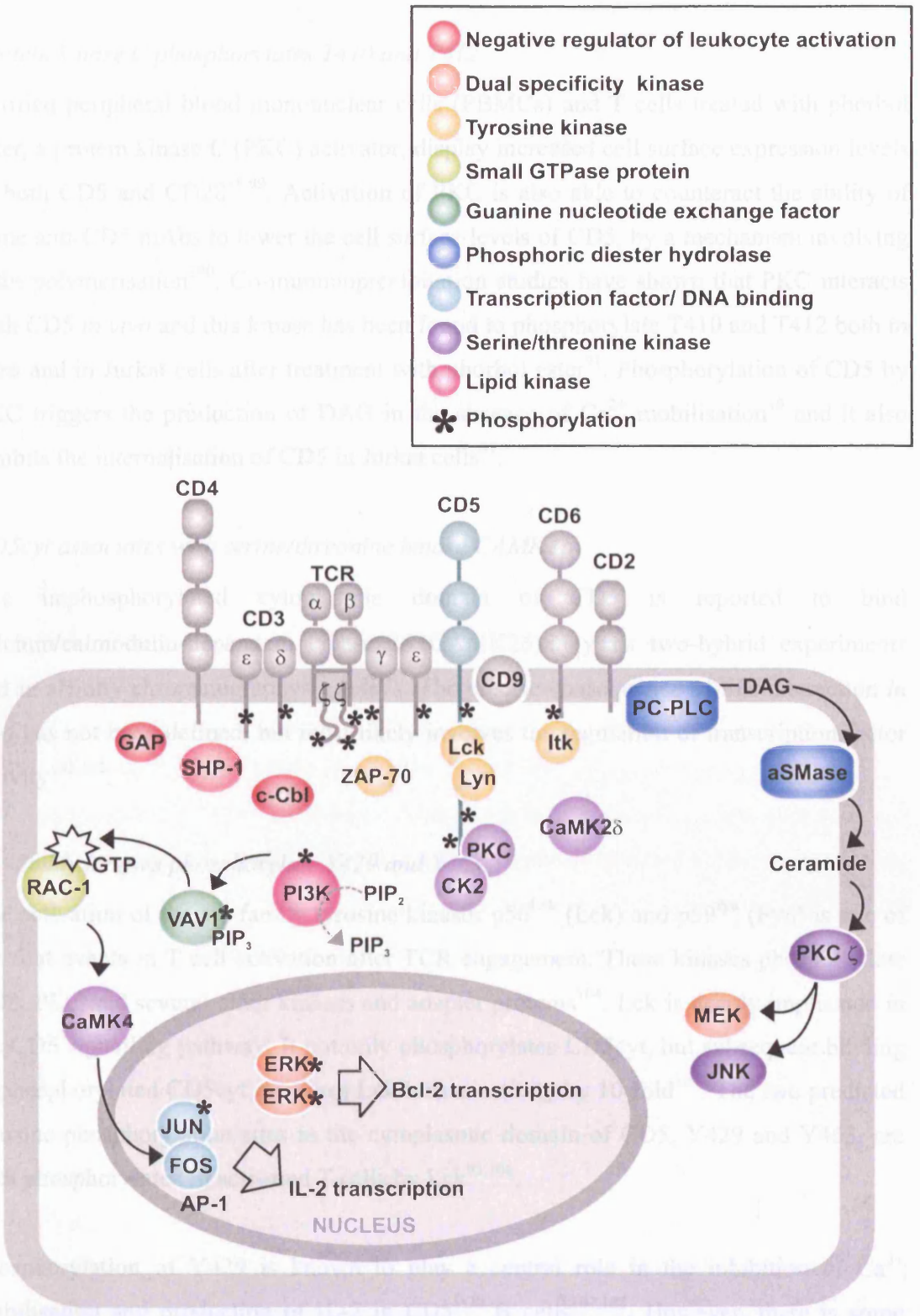


**Figure I.6** A summary of current reports of the protein-protein interactions and signalling pathways involving CD5 in T cells.

*Protein kinase C phosphorylates T410 and T412*

Purified peripheral blood mononuclear cells (PBMCs) and T cells treated with phorbol ester, a protein kinase C (PKC) activator, display increased cell surface expression levels of both CD5 and CD28[98,99]. Activation of PKC is also able to counteract the ability of some anti-CD5 mAbs to lower the cell surface levels of CD5, by a mechanism involving actin polymerisation[100]. Co-immunoprecipitation studies have shown that PKC interacts with CD5 *in vivo* and this kinase has been found to phosphorylate T410 and T412 both *in vitro* and in Jurkat cells after treatment with phorbol ester[91]. Phosphorylation of CD5 by PKC triggers the production of DAG in the absence of $Ca^{2+}$ mobilisation[40] and it also inhibits the internalisation of CD5 in Jurkat cells[91].

*CD5cyt associates with serine/threonine kinase CAMK2δ*

The unphosphorylated cytoplasmic domain of CD5 is reported to bind calcium/calmodulin-dependent kinase 2δ (CAMK2δ) in yeast two-hybrid experiments and in affinity chromatography[101] assays. The precise consequence of this interaction *in vivo* has not been defined, but most likely involves the regulation of transcription factor activity[102,103].

*Src-family kinases phosphorylate Y429 and Y463*

The activation of the src-family tyrosine kinases $p56^{Lck}$ (Lck) and $p59^{fyn}$ (Fyn) is one of the first events in T cell activation after TCR engagement. These kinases phosphorylate CD3, PLC and several other kinases and adapter proteins[104]. Lck is deeply implicated in the CD5 signalling pathway. It not only phosphorylates CD5cyt, but subsequent binding to phosphorylated CD5cyt increases Lck kinase activity by 10-fold[105]. The two predicted tyrosine phosphorylation sites in the cytoplasmic domain of CD5, Y429 and Y463, are both phosphorylated in activated T cells by Lck[93,106].

Phosphorylation of Y429 is known to play a central role in the inhibition of $Ca^{2+}$ mobilisation and production of IL-2 in $CD5^{POS}$ B cells[107,108]. However, there is some disagreement about the identity of kinase responsible for CD5cyt tyrosine phosphorylation in B cells. Lyn kinase was the first to be reported to phosphorylate CD5 in B cells[109]. In contrast to Fyn and Lck, which are mainly considered stimulatory kinases, Lyn is usually implicated in the down-regulation of B cell activation[110,111].

Along with CD5, Lyn has been co-immunoprecipitated with anti-CD5 mAbs from purified B1 cells after stimulation with anti-IgM. Whilst in wild-type B1 cells CD5cyt was found to be tyrosine phosphorylated after anti-IgM stimulation, tyrosine phosphorylation could not be detected in $Lyn^{NEG}$ B1 cells[109].

A recent publication has challenged the function of Lyn as the kinase involved in the phosphorylation of CD5cyt in B cells. The authors of this study propose that this role is fulfilled by Lck, as co-expression of Lck and CD5 was found to confer the hypo-responsive phenotype of B1 cells[68]. $CD5^{NEG}$ B cells do not express Lck[112], but in $CD5^{POS}$ B cells Lck is an important signalling molecule, as Lck activity appears to be necessary for the activation of the MAP kinase cascade triggered by BCR engagement[113].

*CD5cyt associates with tyrosine kinase ZAP-70 via the ζ-chain*

The syk-family kinase ZAP-70 plays an essential role in the activation of T cells. It is recruited to the phosphorylated ζ chain of CD3 where in turn it is phosphorylated by Lck. Phosphorylated ZAP-70 then phosphorylates downstream adapter proteins such as the linker for activation of T cells (LAT) and SH2 domain-containing leukocyte protein (SLP76) (Figure I.3). ZAP-70 has been found by co-immunoprecipitation to be associated with CD5 in freshly isolated thymocytes[24]. However, the interaction does not appear to be direct, as a peptide containing the Tyr phosphorylated CD5cyt region does not bind ZAP-70[114]. The authors suggest that the interaction is via the ζ-chain, because phosphorylated ζ-chain immunoprecipitates, which are able to bind to ZAP-70, also contain CD5[24].

*Phosphorylated CD5cyt could recruit SHP-1*

Phosphatases counter-balance the tyrosine phosphorylation cascade triggered by AgR engagement. One of the most studied phosphatases in AgR signalling is the SH2 domain-bearing protein tyrosine phosphatase 1 (SHP-1). SHP-1 is recruited to the cell membrane following the phosphorylation of specific tyrosine-based motifs in inhibitory receptors. It is phosphorylated and subsequently engages in the dephosphorylation of a plethora of substrates including kinases, adapter proteins and transmembrane receptors. As CD5 is able to dampen lymphocyte activation, it was hypothesised that CD5cyt would recruit SHP-1 and/or other phosphatases in order to initiate down-regulation of AgR signalling.

An association between SHP-1 and phosphorylated CD5 has been detected by co-immunoprecipitation in purified B1 cells[26], mature T cells[61,115] and thymocytes[116]. The phosphotyrosine residue that binds to the SH2 domain of SHP-1 was proposed to be Y378, based on experiments performed in Jurkat T cells with a chimeric construct of the cytoplasmic domain of CD5 and the extracellular domain of CD6[61]. However, this tyrosine is predicted to be at the C-terminal end of the transmembrane region, and other groups have concluded that Y378 is not phosphorylated[93,106]. In B1 cells, the association between SHP-1 and the BCR appears to stabilised by the presence of CD5[26], and there is also evidence that CD5 binding to SHP-1 is mediated by Lyn[109].

The validity of the recruitment of SHP-1 is still debated, as other groups have not been able to detect SHP-1 in association with CD5. It has been reported that in $CD5^{NEG}$ Jurkat T cells expressing a recombinant CD5 transgene, no SHP-1 was detected after stimulation with pervanadate (a tyrosine phosphatase inhibitor) or co-cross-linking with anti-CD3 and anti-CD5 mAbs[93]. In experiments in a B cell line using a chimera of CD5cyt with the extracellular domain of the inhibitory receptor FCγRIIB1, no SHP-1 activity could be found in the anti-FCγRIIB1 immunoprecipitates. However, this chimera was able to antagonise both $Ca^{2+}$ mobilisation and the MAPK pathway, and its inhibitory properties depended on the presence of the CD5cyt region comprising Y429[107].

The role of CD5 in relation with SHP-1 has also been investigated in animal models. $SHP\text{-}1^{NEG}$ mice ('motheaten' mutation, *me*) have severe autoimmune disease and die at three weeks of age of autoimmune pneumonitis. A less severe mutation, $SHP\text{-}1^{LOW}$ (viable motheaten, *mev*) is usually used to study the effects of the absence of SHP-1 because *mev* mice live longer than *me* mice[117]. Both $CD5^{NEG}$ and $SHP\text{-}1^{LOW}$ B cells are hyper-reactive consistent with the hypothesis that CD5 inhibits the AgR signalling pathway by recruitment of SHP-1. According to this premise, $CD5^{NEG}$ *mev* B cells should show even higher hyper-responsiveness than *mev* B cells. Surprisingly however, $CD5^{NEG}$ *mev* mice have increased lifespan and less severe phenotype than $CD5^{POS}$ *mev* mice[117]. The phenotype of $CD5^{NEG}$ *mev* mice illustrates the complexity of the events triggered by gene deletion in the immune system, and perhaps suggest that CD5 is also part of an alternative pathway that is not of an inhibitory nature and does not necessary require SHP-1. This pathway does not have to involve CD5cyt, but could be activated by the extracellular region of CD5 (CD5ext) upon binding to its ligand. Consistent with this

hypothesis, different groups have reported that the application of recombinant CD5ext is able to activate B cells[118-120]. In the absence of SHP-1 CD5cyt would be unable to down-modulate the AgR, which would render cells hyper-reactive, but CD5ext would still be capable of promoting activation via its putative counter-receptor, this would further enhance cell reactivity. Within this model, he loss of CD5 in the absence of SHP-1 would additionally influence the activation pathway triggered by the counter-receptor, thus the cancellation of these positive and negative effects would eliminate the contribution of CD5 to cell reactivity.

*Phosphoinositide 3 kinase binds to phosphorylated CD5*

PI3K is recruited to the membrane once CD3 is phosphorylated upon antigen engagement. Being held close to its substrates, *i.e.* membrane phospholipids, PI3K starts active generation of phosphatidylinositol 3,4,5-trisphosphate ($PIP_3$). $PIP_3$ in turn facilitates recruitment of PLC-γ1 and activates several downstream protein kinases[121].

Association between PI3K and phosphorylated CD5cyt has been documented in activated thymocytes[114], in Jurkat T cells[93] and in purified T cells[122]. The interaction was found to be direct and mediated by the SH2 domains of the p85 adaptor subunit of PI3K. Whilst the C-terminal SH2 domain showed preferential binding to a $pY_{463}$DLH peptide, the N-terminal SH2 domain bound preferentially to a peptide containing pY429 and pY441, indicating that PI3K could potentially bind CD5cyt with both SH2 domains.

*The pathway that leads to the activation of IL-2 gene transcription after CD5 stimulation is different from the pathway triggered by TCR engagement*

$Ca^{2+}$/Calmodulin dependent kinase IV (CaMK IV) is a serine/theonine kinase involved in the up-regulation of the expression of proteins that form the AP-1 transcription factor[123]. CaMK IV also appears to be implicated in decreasing the threshold for positive selection in thymocytes[124]. CaMK IV has been reported to be involved in the expression of IL-2 by purified T cells stimulated with anti-CD5, anti-CD28 and phytohemagglutinin[125]. This pathway of IL-2 expression was reported to involve the phosphorylation of PI3K, Vav, Rac1 and CaMK IV (in that order), and to cause an increase in the DNA binding activity of AP-1 transcription factor, while leaving other transcription factors unchanged[122]. These results suggest that CD5 stimulation is able to trigger an AgR-independent signal transduction cascade that activates one of the most important genes for T cell activation.

*The CD5 signalling cascade in thymocytes might be different to that in mature T cells*

Both cellular-Cbl and the GTPase-activating protein for Ras GTPase (RasGAP), are considered negative regulators of the TCR signalling cascade. The oncoprotein c-Cbl recognises tyrosine kinase substrates and promotes their ligation to ubiquitin marking them for degradation[110,126]. Disruption of the c-Cbl gene causes enhanced T cell responsiveness upon anti-CD3 mAb stimulation[127]. RasGAP is the down-regulator of the small GTPase Ras[128]. Ras activity is necessary for the activation of T cells, as it is up-stream of both the Rac and Raf pathways. In isolated thymocytes stimulated with pervanadate and cross-linked with anti-CD5 Abs, CD5cyt associates with both c-Cbl and rasGAP[129]. However, none of these associations was found in mature T cells[93], suggesting that this could be a reason for the different effects of CD5 in T cells and thymocytes.

It has been reported that in DP thymocytes, cross-linking of CD5 and CD3 causes a reduction in the levels of intracellular $Ca^{2+}$, as expected for an inhibitor of the TCR, but an increased level of activity for the MAP kinase cascade component ERK[56]. This increased ERK activity stimulates the production of B cell lymphoma protein 2 (Bcl-2)[56], an oncoprotein that promotes cell survival by inhibiting apoptosis[130].

*An undetermined ceramide-mediated pathway becomes active after CD5 ligation*

CD5 has been found to be involved in an undefined pathway that uses ceramide as a second messenger. Stimulation of T cells with anti-CD5, anti-CD3 and anti-CD28 mAbs activates two phosphoric diester hydrolases: acidic sphingomyelinase (aSMase)[131] and phosphatidylcholine-specific phospholipase C (PC-PLC)[58]. These enzymes catalyse the formation of ceramide and 1,2-diacyl-glycerols (DAG) from the cell membrane components sphingomyelin and phosphatidylcholine respectively[132-134]. Activation of aSMase was severely compromised by mutation of the Ser residues phosphorylated by CK2[131], which could mean that CK2 is also implicated in this pathway. The group also found that other components of this signalling cascade are the ceramide-activated protein kinase C-ζ (PKC-ζ), mitogen activated protein kinase kinase (MEK) and c-Jun $NH_2$-terminal kinase (JNK), but the final outcome of the pathway is yet to be elucidated.

## Interactions with other extracellular proteins

As well as being in close contact with the AgRs and CD4 or CD8[135], CD5 has been reported to be in physical proximity with three other leukocyte cell surface transmembrane proteins, CD2, CD6 and CD9. The interaction between CD5 and CD9 in T cells was discovered in co-immunoprecipitation and capping experiments. CD9 does not have a cytoplasmic domain for signal transduction, and it is thought to function in association with other molecules[136]. CD9 belongs to a family of transmembrane proteins that function as 'molecular facilitators' and its interaction is most likely to assist the distribution of CD5 on the cell surface[137].

CD2 is a lymphocyte receptor involved in the stabilisation of the IS and in signal transduction[138]. The binding of CD2 with its ligand (CD58 in humans and CD48 in rats) aids T cell activation[139] and contributes to the adhesion to the APC[140]. The direct interaction of CD2 with CD5 was detected by co-immunoprecipitation and co-capping in T cells and thymocytes[141]. Both the extracellular and the cytoplasmic domain of CD2 appear to be involved in the association[141]. Stimulation of CD2 with anti-CD2 Abs was reported to induce dephosphorylation of CD5 in T cells, presumably by recruitment of SHP-1[115]. $CD5^{NEG}$ and $CD2^{NEG}$ mice have similar phenotypes, in that $CD2^{NEG}$ mice seem healthy and able to mount a strong immune response[142]. As $CD5^{NEG}$ thymocytes, $CD2^{NEG}$ thymocytes have enhanced negative and positive selection[143], indicating that CD2 is inhibiting TCR signalling events in thymocytes, the opposite effect than observed in mature T cells stimulated with anti-CD2 mAbs[144].

CD5 and CD6 are thought to have evolved from a gene duplication event[145,146]. They have very similar domain organisation and are expressed by the same cell types[147]. The function of CD6 in lymphocyte activation is not well established but, as with CD5, it is known to be involved in the modulation of lymphocyte activation and thymocyte development. CD6 is thought to be an enhancer of the AgR[148] and to favour positive selection in thymocytes[149]. Antibodies anti-CD6 are able to activate resting T cells[150-152]. It has also been reported that CD6 is able to rescue B-CLL cells from apoptosis by up-regulating the expression of the oncoprotein Bcl-2[153] and that it enhances T cell adhesion to the APC[154]. FRET, co-immunoprecipitation and co-capping experiments have shown that CD6, the same as CD5, associates with the TCR[154]. CD6 and CD5 are able to co-immunoprecipitate[155] in T cells, and their physical proximity in resting and activated T

cells has been detected by FRET[156]. Stimulation of CD6 with anti-CD6 Abs has been reported to increase the tyrosine phosphorylation level of CD5[155]. CD6 is able to recruit Lck and also the IL-2 inducible T cell kinase (Itk). Itk is critical for PLC-γ1 activation and it does not appear to be associated with CD5 or the TCR[155]. A combination of Lck and Itk is able to efficiently phosphorylate a CD5cyt peptide containing Tyr 429 and Tyr441, suggesting that Itk may also be a part of the signalling pathway of CD5[155].

The reported interactions with CD6 and CD2 show that the nature of the effects of CD5 on the modulation of lymphocyte activation is more complex than has been foreseen. These observations raise the possibility that the activation of CD2 and/or CD6 is able influence the outcome of the transduction pathways associated with CD5. Perhaps these receptors have redundant functions, which would explain why $CD5^{NEG}$ and $CD2^{NEG}$ mice do not possess a very distinctive phenotype, and why the extracellular domain of CD5 appears to be dispensable for thymocyte development.

## The extracellular ligand of CD5

The identification of the molecular ligand(s) for extracellular region of CD5 has been controversial as reported interactions have proved difficult to reproduce. Furthermore, there have not been any recent publications concerning most of the proposed ligands, possibly indicating that the target molecules remain elusive. Most of the identified potential ligands (see below) are able to activate B cells[118,120,157-159]. Although there is always a possibility that CD5 might not have a ligand, the fact that the administration of soluble forms of CD5 ameliorates certain experimental autoimmune diseases *in vivo*[159,160], that soluble CD5 can promote B cell proliferation[120], and the very existence of a natural soluble form of CD5[161], suggest that there is a lymphocyte-derived molecule capable of binding CD5ext.

The first proposed ligand for CD5 was CD72[162], a B cell C-type lectin (a $Ca^{2+}$-dependent carbohydrate recognition receptor). The interaction was discovered by incubating purified biotinylated CD5 with several cell lineages. Subsequently, it was also reported that purified CD5 binds immobilised recombinant CD72 in an enzyme-linked immunosorbent assay (ELISA)[118]. In both cases the binding of CD5 to the target could only be prevented by addition of anti-CD72[163]. However, no independent research has been able to reproduce these findings[120,159,164,165].

Two inducible unidentified glycoproteins, present in both activated B and T cells, have been reported to be able to bind soluble recombinant constructs of CD5. Glycoproteins gp40[120,159] and gp80[120] were pulled-down from activated lymphocyte cell lysates using recombinant CD5 as a bait[120,159]. This interaction was found to be divalent metal cation-dependent and to depend on the glycosylation of CD5, which suggests that the ligand is a C-type lectin, but it was actively confirmed that the receptor was not CD72[159] in this case. A similar association was reported independently: a soluble recombinant construct of bovine CD5 was found to interact with a 200 kDa protein in activated B cells[158].

A broadly-distributed constitutively-expressed glycoprotein, gp150, has also been proposed to be a ligand for CD5. As opposed to the putative C-lectin ligands, the binding of gp150 to a recombinant form of CD5 was independent of divalent metal cations and was only disrupted at high concentrations of monosaccharides[165]. The interaction was discovered by incubating several cell lines with a recombinant soluble CD5. Gp150 could not be cross-linked with CD5 under the conditions that were attempted, but could be isolated by co-immunoprecipitation with anti-CD5 mAbs[165].

CD5 has been reported to bind the immunoglobulin domain that constitutes the variable region of the heavy chain ($V_H$) of the BCR. The binding has been mapped to the second SRCR domain of CD5 (CD5d2) and framework region of the $V_H$[119,166] (the face of the domain that presents lower variability within different $V_H$ domains compared to the CDRs). Purified recombinant CD5, and CD5 from lysates of several kinds of B cells can be immunoprecipitated using purified immobilised Ig or Ab fragments. The binding is not specific to the kind of Ab used, arguing that it does not depend on Ab specificity[119]. If this is true, all experiments on lymphocytes involving the use of antibodies could influence the status of CD5. It was proposed that this interaction is involved in the long-

term survival of peripheral B cells by delivering mild stimulatory signals, just as is claimed for low-affinity self-antigens[167].

## Origin and scope of the thesis

The interest in determining the 3D structure of the extracellular domains of CD5 arose more than 8 years ago from a collaboration with Prof Neil Barclay and Dr Marion Brown at the Sir William Dunn School of Pathology in Oxford. There were two strong motivations for attempting to obtain the structure of the SRCR domains of CD5: to enable a structure-function analysis of CD5; and to establish if SRCR domains were members of the immunoglobulin superfamily or not, as was debated at that time. The N-terminal domain of CD5 (CD5d1) was chosen as the primary target for structural investigation because it was thought most likely to contain the ligand-binding site as it is the most distant from the cell surface. Furthermore, the presence of the stalk region between CD5d1 and CD5d2 suggested that it was unlikely to require the other domains to achieve its native fold.

CD5d1 was expressed in Chinese hamster ovary cells in sufficient quantities to establish by $^{1}$H-NMR that the domain was folded and globular. Because NMR structure determination of a 15 kDa protein requires several milligrams of $^{15}$N and $^{13}$C isotope-labelled protein, efforts were made to develop a yeast expression system. However, the levels of expression in the yeast *Pichia pastoris* were always low and labelling with $^{13}$C proved prohibitively expensive. Using glycosydase-treated $^{15}$N single-labelled samples, it was possible to assign *ca.* 85% of the backbone amide resonances, but further progress was hampered by substantial peak overlap.

Inspired by the successes of other laboratories in expressing unglycosylated versions of glycoproteins in *Escherichia coli*, the bacterial expression of CD5d1 was then pursued. The bacterial CD5d1 was found to be expressed in an insoluble form in inclusion bodies. *In vitro* folding by dilution was attempted with very limited yields and poor reproducibility. However, sufficient $^{15}$N-labelled sample was produced to confirm that the unglycosylated construct was folded and that the NMR spectra of the bacterial CD5d1 were similar to those of the yeast and mammalian versions, implying that the N-glycan was not essential for proper folding. Such was the state of affairs when I took

over the project.

In order to proceed with the resonance assignment and progress towards the determination of the 3D solution structure of CD5d1, further optimisation of the *in vitro* folding methodology and improvement of protein solubility were essential. This thesis describes the strategies that I developed to produce correctly folded double isotope-labelled CD5d1 at suitable concentrations for three-dimensional NMR experiments. The NMR data recorded with these samples has enabled us to determine the long-sought solution structure of CD5d1. The NMR structure of CD5d1 presented here provides a framework for analysis of the structure-function characteristics of both CD5 in particular and SRCR domain-containing proteins in general.

The approaches that I have developed to optimise *in vitro* folding conditions and improve solubility are not only applicable to CD5d1, but are general methods that could be applied, straightforwardly, to other proteins. In particular, the application of these methods will allow the future 3D structure determination of the other SRCR domains of CD5 and CD6, which will provide additional information for the future understanding of the interactions of CD5 and CD6 with their putative ligands.

## References


1. Hasler P & Zouali M (2001) B cell receptor signaling and autoimmunity. *FASEB J.* **15**, 2085-2098.
2. Love PE & Chan AC (2003) Regulation of thymocyte development: only the meek survive. *Curr Opin Immunol.* **15**, 199-203.
3. Zamoyska R & Lovatt M (2004) Signalling in T-lymphocyte development: integration of signalling pathways is the key. *Curr Opin Immunol.* **16**, 191-196.
4. Killeen N, Irving BA, Pippig S & Zingler K (1998) Signaling checkpoints during the development of T lymphocytes. *Curr Opin Immunol.* **10**, 360-367.
5. Cannons JL & Schwartzberg PL (2004) Fine-tuning lymphocyte regulation: what's new with tyrosine kinases and phosphatases? *Curr Opin Immunol.* **16**, 296-303.
6. Weiss A & Cambier JC (2004) Lymphocyte activation. *Curr Opin Immunol.* **16**, 285-287.
7. Macian F, Lopez-Rodriguez C & Rao A (2001) Partners in transcription: NFAT and AP-1. *Oncogene.* **20**, 2476-2489.
8. Schmitz ML, Bacher S & Dienz O (2003) NF-kappaB activation pathways induced by T cell costimulation. *FASEB J.* **17**, 2187-2193.
9. Yordy JS & Muise-Helmericks RC (2000) Signal transduction and the Ets family of transcription

factors. *Oncogene* . **19**, 6503-6513.

10. Leibson PJ (2004) The regulation of lymphocyte activation by inhibitory receptors. *Curr Opin Immunol.* **16**, 328-336.
11. Azzam HS, Grinberg A, Lui K, Shen H, Shores EW & Love PE (1998) CD5 expression is developmentally regulated by T cell receptor (TCR) signals and TCR avidity. *J Exp Med.* **188**, 2301-2311.
12. Kaplan D, Smith D, Meyerson H, Pecora N & Lewandowska K (2001) CD5 expression by B lymphocytes and its regulation upon Epstein-Barr virus transformation. *Proc Natl Acad Sci USA.* **98**, 13850-13853.
13. Raman C & Knight KL (1992) CD5+ B cells predominate in peripheral tissues of rabbit. *J Immunol.* **149**, 3858-3864.
14. Koskinen R, Gobel TW, Tregaskes CA, Young JR & Vainio O (1998) The structure of avian CD5 implies a conserved function. *J Immunol.* **160**, 4943-4950.
15. Berland R & Wortis HH (1998) An NFAT-dependent enhancer is necessary for anti-IgM-mediated induction of murine CD5 expression in primary splenic B cells. *J Immunol.* **161**, 277-285.
16. Kearse KP, Takahama Y, Punt JA, Sharrow SO & Singer A (1995) Early molecular events induced by T cell receptor (TCR) signaling in immature CD4+ CD8+ thymocytes: increased synthesis of TCR-alpha protein is an early response to TCR signaling that compensates for TCR-alpha instability, improves TCR assembly, and parallels other indicators of positive selection. *J Exp Med.* **181**, 193-202.
17. Dutz JP, Ong CJ, Marth J & Teh HS (1995) Distinct differentiative stages of CD4+CD8+ thymocyte development defined by the lack of coreceptor binding in positive selection. *J Immunol.* **154**, 2588-2599.
18. Azzam HS, DeJarnette JB, Huang K, Emmons R, Park CS, Sommers CL, El Khoury D, Shores EW & Love PE (2001) Fine tuning of TCR signaling by CD5. *J Immunol.* **166**, 5464-5472.
19. Berland R & Wortis HH (2002) Origins and functions of B-1 cells with notes on the role of CD5. *Annu Rev Immunol.* **20**, 253-300.
20. Yang Y, Contag CH, Felsher D, Shachaf CM, Cao Y, Herzenberg LA, Herzenberg LA & Tung JW (2004) The E47 transcription factor negatively regulates CD5 expression during thymocyte development. *Proc Natl Acad Sci USA.* **101**, 3898-3902.
21. Brossard C, Semichon M, Trautmann A & Bismuth G (2003) CD5 inhibits signaling at the immunological synapse without impairing its formation. *J Immunol.* **170**, 4623-4629.
22. Osman N, Lazarovits AI & Crumpton MJ (1993) Physical association of CD5 and the T cell receptor/CD3 antigen complex on the surface of human T lymphocytes. *Eur J Immunol.* **23**, 1173-1176.
23. Burgess KE, Yamamoto M, Prasad KV & Rudd CE (1992) CD5 acts as a tyrosine kinase substrate within a receptor complex comprising T-cell receptor zeta chain/CD3 and protein-tyrosine kinases p56lck and p59fyn. *Proc Natl Acad Sci USA.* **89**, 9311-9315.
24. Gary-Gouy H, Lang V, Sarun S, Boumsell L & Bismuth G (1997) In vivo association of CD5 with tyrosine-phosphorylated ZAP-70 and p21 phospho-zeta molecules in human CD3+ thymocytes. *J Immunol.* **159**, 3739-3747.

25. Lankester AC, van Schijndel GM, Cordell JL, van Noesel CJ & van Lier RA (1994) CD5 is associated with the human B cell antigen receptor complex. *Eur J Immunol.* **24**, 812-816.
26. Sen G, Bikah G, Venkataraman C & Bondada S (1999) Negative regulation of antigen receptor-mediated signaling by constitutive association of CD5 with the SHP-1 protein tyrosine phosphatase in B-1 B cells. *Eur J Immunol.* **29**, 3319-3328.
27. Jurgens JB, Gartland LA, Du PL, Horton JD, Gobel TW & Cooper MD (1995) Identification of a candidate CD5 homologue in the amphibian *Xenopus laevis*. *J Immunol.* **155**, 4218-4223.
28. Tarakhovsky A, Muller W & Rajewsky K (1994) Lymphocyte populations and immune responses in CD5-deficient mice. *Eur J Immunol.* **24**, 1678-1684.
29. Kissler S, Lu L & Cantor H (2004) Thymic selection can compensate for mutations affecting T cell activation and generate a normal T cell repertoire in mutant mice. *Proc Natl Acad Sci USA.* **101**, 210-214.
30. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* . **292**, 195-202.
31. Julenius K, Molgaard A, Gupta R & Brunak S (2005) Prediction, conservation analysis and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology.* **15**,153-164
32. McAlister MS, Brown MH, Willis AC, Rudd PM, Harvey DJ, Aplin R, Shotton DM, Dwek RA, Barclay AN & Driscoll PC (1998) Structural analysis of the CD5 antigen -expression, disulphide bond analysis and physical characterisation of CD5 scavenger receptor superfamily domain 1. *Eur J Biochem.* **257**, 131-141.
33. Rudd PM, Wormald MR, Harvey DJ, Devasahayam M, McAlister MS, Brown MH, Davis SJ, Barclay AN & Dwek RA (1999) Oligosaccharide analysis and molecular modeling of soluble forms of glycoproteins belonging to the Ly-6, scavenger receptor, and immunoglobulin superfamilies expressed in Chinese hamster ovary cells. *Glycobiology.* **9** , 443-458.
34. Brown MH & Barclay AN (1994) Expression of immunoglobulin and scavenger receptor superfamily domains as chimeric proteins with domains 3 and 4 of CD4 for ligand analysis. *Protein Eng.* **7**, 515-521.
35. Starling GC, Llewellyn MB, Whitney GS & Aruffo A (1997) The Ly-1.1 and Ly-1.2 epitopes of murine CD5 map to the membrane distal scavenger receptor cysteine-rich domain. *Tissue Antigens.* **49**, 1-6.
36. Calvo J, Padilla O, Places L, Vigorito E, Vila JM, Vilella R, Mila J, Vives J, Bowen MA & Lozano F (1999) Relevance of individual CD5 extracellular domains on antibody recognition, glycosylation and co-mitogenic signalling. *Tissue Antigens.* **54**, 16-26.
37. Lozano F, Simarro M, Calvo J, Vila JM, Padilla O, Bowen MA & Campbell KS (2000) CD5 signal transduction: positive or negative modulation of antigen receptor signaling. *Crit Rev Immunol.* **20**, 347-358.
38. Youinou P, Jamin C & Lydyard PM (1999) CD5 expression in human B-cell populations. *Immunol Today.* **20**, 312-316.
39. Raman C (2002) CD5, an important regulator of lymphocyte selection and immune tolerance. *Immunol Res.* **26**, 255-263.

40. Alberola-Ila J, Places L, Cantrell DA, Vives J & Lozano F (1992) Intracellular events involved in CD5-induced human T cell activation and proliferation. *J Immunol.* **148**, 1287-1293.
41. Spertini F, Stohl W, Ramesh N, Moody C & Geha RS (1991) Induction of human T cell proliferation by a monoclonal antibody to CD5. *J Immunol.* **146**, 47-52.
42. Berney SM, Schaan T, Wolf RE, Kimpel DL, van der HH & Atkinson TP (2001) CD5 (OKT1) augments CD3-mediated intracellular signaling events in human T lymphocytes. *Inflammation.* **25**, 215-221.
43. Bikah G, Carey J, Ciallella JR, Tarakhovsky A & Bondada S (1996) CD5-mediated negative regulation of antigen receptor-induced growth signals in B-1 B cells. *Science.* **274**, 1906-1909.
44. Bondada S, Bikah G, Robertson DA & Sen G (2000) Role of CD5 in growth regulation of B-1 cells. *Curr Top Microbiol Immunol.* **252:141-9**, 141-149.
45. Smith K, Seddon B, Purbhoo MA, Zamoyska R, Fisher AG & Merkenschlager M (2001) Sensory adaptation in naive peripheral CD4 T cells. *J Exp Med.* **194**, 1253-1261.
46. Stamou P, de Jersey J, Carmignac D, Mamalaki C, Kioussis D & Stockinger B (2003) Chronic exposure to low levels of antigen in the periphery causes reversible functional impairment correlating with changes in CD5 levels in monoclonal CD8 T cells. *J Immunol.* **171**, 1278-1284.
47. Wong P, Barton GM, Forbush KA & Rudensky AY (2001) Dynamic tuning of T cell reactivity by self-peptide-major histocompatibility complex ligands. *J Exp Med.* **193**, 1179-1187.
48. Kieper WC, Burghardt JT & Surh CD (2004) A role for TCR affinity in regulating naive T cell homeostasis. *J Immunol.* **172**, 40-44.
49. Neuberger MS (1997) Antigen receptor signaling gives lymphocytes a long life. *Cell.* **90**, 971-973.
50. Ledbetter JA, Parsons M, Martin PJ, Hansen JA, Rabinovitch PS & June CH (1986) Antibody binding to CD5 (Tp67) and Tp44 T cell surface molecules: effects on cyclic nucleotides, cytoplasmic free calcium, and cAMP-mediated suppression. *J Immunol.* **137**, 3299-3305.
51. McAteer MJ, Lagarde AC, Georgiou HM & Bellgrau D (1988) A requirement for the CD5 antigen in T cell activation. *Eur J Immunol.* **18**, 1111-1117.
52. Kroesen BJ, Bakker A, van Lier RA, The HT & de Leij L (1995) Bispecific antibody-mediated target cell-specific costimulation of resting T cells via CD5 and CD28. *Cancer Res.* **55**, 4409-4415.
53. Ceuppens JL & Baroja ML (1986) Monoclonal antibodies to the CD5 antigen can provide the necessary second signal for activation of isolated resting T cells by solid-phase-bound OKT3. *J Immunol.* **137**, 1816-1821.
54. June CH, Rabinovitch PS & Ledbetter JA (1987) CD5 antibodies increase intracellular ionized calcium concentration in T cells. *J Immunol.* **138**, 2782-2792.
55. Imboden JB, June CH, McCutcheon MA & Ledbetter JA (1990) Stimulation of CD5 enhances signal transduction by the T cell antigen receptor. *J Clin Invest.* **85**, 130-134.
56. Zhou XY, Yashiro-Ohtani Y, Toyo-oka K, Park CS, Tai XG, Hamaoka T & Fujiwara H (2000) CD5 costimulation up-regulates the signaling to extracellular signal-regulated kinase activation in CD4+CD8+ thymocytes and supports their differentiation to the CD4 lineage. *J Immunol* . **164**, 1260-1268.

57. Verwilghen J, Vandenberghe P, Wallays G, de Boer M, Anthony N, Panayi GS & Ceuppens JL (1993) Simultaneous ligation of CD5 and CD28 on resting T lymphocytes induces T cell activation in the absence of T cell receptor/CD3 occupancy. *J Immunol.* **150**, 835-846.

58. Simarro M, Pelassy C, Calvo J, Places L, Aussel C & Lozano F (1997) The cytoplasmic domain of CD5 mediates both TCR/CD3-dependent and -independent diacylglycerol production. *J Immunol.* **159**, 4307-4315.

59. Tarakhovsky A, Kanner SB, Hombach J, Ledbetter JA, Muller W, Killeen N & Rajewsky K (1995) A role for CD5 in TCR-mediated signal transduction and thymocyte selection. *Science.* **269**, 535-537.

60. Pena-Rossi C, Zuckerman LA, Strong J, Kwan J, Ferris W, Chan S, Tarakhovsky A, Beyers AD & Killeen N (1999) Negative regulation of CD4 lineage development and responses by CD5. *J Immunol.* **163**, 6494-6501.

61. Perez-Villar JJ, Whitney GS, Bowen MA, Hewgill DH, Aruffo AA & Kanner SB (1999) CD5 negatively regulates the T-cell antigen receptor signal transduction pathway: involvement of SH2-containing phosphotyrosine phosphatase SHP-1. *Mol Cell Biol.* **19** , 2903-2912.

62. Jung S (2004) Good, bad and beautiful--the role of dendritic cells in autoimmunity. *Autoimmun Rev.* **3**, 54-60.

63. Hawiger D, Masilamani RF, Bettelli E, Kuchroo VK & Nussenzweig MC (2004) Immunological unresponsiveness characterized by increased expression of CD5 on peripheral T cells induced by dendritic cells in vivo. *Immunity.* **20**, 695-705.

64. Bhandoola A, Bosselut R, Yu Q, Cowan ML, Feigenbaum L, Love PE & Singer A (2002) CD5-mediated inhibition of TCR signaling during intrathymic selection and development does not require the CD5 extracellular domain. *Eur J Immunol.* **32**, 1811-1817.

65. Page DM (1999) Cutting edge: thymic selection and autoreactivity are regulated by multiple coreceptors involved in T cell activation. *J Immunol.* **163**, 3577-3581.

66. Kishimoto H & Sprent J (1999) Several different cell surface molecules control negative selection of medullary thymocytes. *J Exp Med.* **190**, 65-73.

67. Boes M (2000) Role of natural and immune IgM antibodies in immune responses. *Mol Immunol.* **37**, 1141-1149.

68. Dal Porto JM, Burke K & Cambier JC (2004) Regulation of BCR signal transduction in B-1 cells requires the expression of the Src family kinase Lck. *Immunity.* **21**, 443-453.

69. Stall AM, Wells SM & Lam KP (1996) B-1 cells: unique origins and functions. *Semin Immunol.* **8**, 45-59.

70. Dono M, Cerruti G & Zupo S (2004) The CD5+ B-cell. *Int J Biochem Cell Biol.* **36**, 2105-2111.

71. Hayakawa K & Hardy RR (2000) Development and function of B-1 cells. *Curr Opin Immunol.* **12**, 346-353.

72. Popi AF, Lopes JD & Mariano M (2004) Interleukin-10 secreted by B-1 cells modulates the phagocytic activity of murine macrophages in vitro. *Immunology.* **113**, 348-354.

73. Paciorkowski N, Porte P, Shultz LD & Rajan TV (2000) B1 B lymphocytes play a critical role in host protection against lymphatic filarial parasites. *J Exp Med.* **191**, 731-736.

74. Youinou P, Pers JO, Jamin C & Lydyard PM (2000) CD5-positive B cells at the crossroads of B cell malignancy and nonorgan-specific autoimmunity. *Pathol Biol (Paris).* **48**, 574-576.

75. Barista I, Romaguera JE & Cabanillas F (2001) Mantle-cell lymphoma. *Lancet Oncol.* **2**, 141-148.

76. Caligaris-Cappio F (1996) B-chronic lymphocytic leukemia: a malignancy of anti-self B cells. *Blood.* **87**, 2615-2620.

77. Pers JO, Jamin C, Le Corre R, Lydyard PM & Youinou P (1998) Ligation of CD5 on resting B cells, but not on resting T cells, results in apoptosis. *Eur J Immunol.* **28**, 4170-4176.

78. Morris DL & Rothstein TL (1994) Decreased surface IgM receptor-mediated activation of phospholipase C gamma 2 in B-1 lymphocytes. *Int Immunol.* **6**, 1011-1016.

79. Wong SC, Chew WK, Tan JE, Melendez AJ, Francis F & Lam KP (2002) Peritoneal CD5+ B-1 cells have signaling properties similar to tolerant B cells. *J Biol Chem.* **277**, 30707-30715.

80. Hippen KL, Tze LE & Behrens TW (2000) CD5 maintains tolerance in anergic B cells. *J Exp Med.* **191**, 883-890.

81. Ishida H, Hastings R, Kearney J & Howard M (1992) Continuous anti-interleukin 10 antibody administration depletes mice of Ly-1 B cells but not conventional B cells. *J Exp Med.* **175**, 1213-1220.

82. Gary-Gouy H, Harriague J, Bismuth G, Platzer C, Schmitt C & Dalloul AH (2002) Human CD5 promotes B-cell survival through stimulation of autocrine IL-10 production. *Blood.* **100**, 4537-4543.

83. Gagro A, McCloskey N, Challa A, Holder M, Grafton G, Pound JD & Gordon J (2000) CD5-positive and CD5-negative human B cells converge to an indistinguishable population on signalling through B-cell receptors and CD40. *Immunology.* **101**, 201-209.

84. Cong YZ, Rabin E & Wortis HH (1991) Treatment of murine CD5- B cells with anti-Ig, but not LPS, induces surface CD5: two B-cell activation pathways. *Int Immunol.* **3**, 467-476.

85. Weichert TR & Schwartz RC (1997) Induction of murine CD5 expression by v-H-ras. *Immunology.* **90**, 30-37.

86. Rothstein TL (2002) Cutting edge commentary: two B-1 or not to be one. *J Immunol.* **168**, 4257-4261.

87. Nakayama K, Takahashi K, Shultz LD, Miyakawa K & Tomita K (1997) Abnormal development and differentiation of macrophages and dendritic cells in viable motheaten mutant mice deficient in haematopoietic cell phosphatase. *Int J Exp Pathol.* **78**, 245-257.

88. Borrello MA, Palis J & Phipps RP (2001) The relationship of CD5+ B lymphocytes to macrophages: insights from normal biphenotypic B/macrophage cells. *Int Rev Immunol.* **20**, 137-155.

89. Blom N, Gammeltoft S & Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol.* **294**, 1351-1362.

90. Alberola-Ila J, Places L, Lozano F & Vives J (1993) Association of an activation inducible serine kinase activity with CD5. *J Immunol.* **151**, 4423-4430.

91. Vila JM, Calvo J, Places L, Padilla O, Arman M, Gimferrer I, Aussel C, Vives J & Lozano F (2001) Role of two conserved cytoplasmic threonine residues (T410 and T412) in CD5 signaling. *J Immunol.* **166**, 396-402.

92. Davies AA, Ley SC & Crumpton MJ (1992) CD5 is phosphorylated on tyrosine after stimulation of the T-cell antigen receptor complex. *Proc Natl Acad Sci USA.* **89**, 6368-6372.

93. Dennehy KM, Ferris WF, Veenstra H, Zuckerman LA, Killeen N & Beyers AD (2001) Determination of the tyrosine phosphorylation sites in the T cell transmembrane glycoprotein CD5. *Int Immunol.* **13**, 149-156.

94. Calvo J, Vilda JM, Places L, Simarro M, Padilla O, Andreu D, Campbell KS, Aussel C & Lozano F (1998) Human CD5 signaling and constitutive phosphorylation of C-terminal serine residues by casein kinase II. *J Immunol.* **161**, 6022-6029.

95. Raman C, Kuo A, Deshane J, Litchfield DW & Kimberly RP (1998) Regulation of casein kinase 2 by direct interaction with cell surface receptor CD5. *J Biol Chem.* **273**, 19183-19189.

96. Meggio F & Pinna LA (2003) One-thousand-and-one substrates of protein kinase CK2? *FASEB J.* **17**, 349-368.

97. Lu X, Axtell RC, Collawn JF, Gibson A, Justement LB & Raman C (2002) AP2 adaptor complex-dependent internalization of CD5: differential regulation in T and B cells. *J Immunol.* **168**, 5612-5620.

98. Carrera AC, Cardenas L, Tugores A, Alonso M, Sanchez-Madrid F & de Landazuri MO (1989) Activators of protein kinase C up-regulate the cell surface expression of CD2 and CD5 T cell glycoproteins. *J Biol Chem.* **264**, 15650-15655.

99. Lozano F, Alberola-Ila J, Places L, Gallart T & Vives J (1990) Protein kinase C-dependent up-regulation of CD5 surface expression on normal and lymphoblastoid T cells. *Immunology.* **70**, 434-439.

100. Alberola-Ila J, Places L, Fabregat V, Vives J & Lozano F (1993) Different mechanisms regulate the monoclonal antibody-induced modulation of CD2, CD3, and CD5 in human lymphocytes. *Cell Immunol.* **147**, 247-255.

101. Bauch A, Campbell KS & Reth M (1998) Interaction of the CD5 cytoplasmic domain with the $Ca^{2+}$/calmodulin-dependent kinase IIdelta. *Eur J Immunol.* **28**, 2167-2177.

102. Hughes K, Edin S, Antonsson A & Grundstrom T (2001) Calmodulin-dependent kinase II mediates T cell receptor/CD3- and phorbol ester-induced activation of IkappaB kinase. *J Biol Chem.* **276**, 36008-36013.

103. Liu H & Grundstrom T (2002) Calcium regulation of GM-CSF by calmodulin-dependent kinase II phosphorylation of Ets1. *Mol Biol Cell.* **13**, 4497-4507.

104. Zamoyska R, Basson A, Filby A, Legname G, Lovatt M & Seddon B (2003) The influence of the src-family kinases, Lck and Fyn, on T cell differentiation, survival and activation. *Immunol Rev.* **191**, 107-118.

105. Raab M, Yamamoto M & Rudd CE (1994) The T-cell antigen CD5 acts as a receptor and substrate for the protein-tyrosine kinase p56lck. *Mol Cell Biol.* **14**, 2862-2870.

106. Vila JM, Gimferrer I, Padilla O, Arman M, Places L, Simarro M, Vives J & Lozano F (2001) Residues Y429 and Y463 of the human CD5 are targeted by protein tyrosine kinases. *Eur J Immunol.* **31**, 1191-1198.

107. Gary-Gouy H, Bruhns P, Schmitt C, Dalloul A, Daeron M & Bismuth G (2000) The pseudo-immunoreceptor tyrosine-based activation motif of CD5 mediates its inhibitory action on B-cell receptor signaling. *J Biol Chem.* **275**, 548-556.

108. Gary-Gouy H, Harriague J, Dalloul A, Donnadieu E & Bismuth G (2002) CD5-negative regulation of B cell receptor signaling pathways originates from tyrosine residue Y429 outside an immunoreceptor tyrosine-based inhibitory motif. *J Immunol.* **168**, 232-239.

109. Ochi H & Watanabe T (2000) Negative regulation of B cell receptor-mediated signaling in B-1 cells through CD5 and Ly49 co-receptors via Lyn kinase activity. *Int Immunol.* **12**, 1417-1423.

110. Latour S & Veillette A (2001) Proximal protein tyrosine kinases in immunoreceptor signaling. *Curr Opin Immunol.* **13**, 299-306.

111. DeFranco AL, Chan VW & Lowell CA (1998) Positive and negative roles of the tyrosine kinase Lyn in B cell function. *Semin Immunol.* **10**, 299-307.

112. Majolini MB, D'Elios MM, Galieni P, Boncristiano M, Lauria F, Del Prete G, Telford JL & Baldari CT (1998) Expression of the T-cell-specific tyrosine kinase Lck in normal B-1 cells and in chronic lymphocytic leukemia B cells. *Blood.* **91**, 3390-3396.

113. Ulivieri C, Valensin S, Majolini MB, Matthews RJ & Baldari CT (2003) Normal B-1 cell development but defective BCR signaling in Lck-/- mice. *Eur J Immunol.* **33**, 441-445.

114. Dennehy KM, Broszeit R, Garnett D, Durrheim GA, Spruyt LL & Beyers AD (1997) Thymocyte activation induces the association of phosphatidylinositol 3-kinase and pp120 with CD5. *Eur J Immunol.* **27**, 679-686.

115. Carmo AM, Castro MA & Arosa FA (1999) CD2 and CD3 associate independently with CD5 and differentially regulate signaling through CD5 in Jurkat T cells. *J Immunol.* **163**, 4238-4245.

116. Pani G, Fischer KD, Mlinaric-Rascan I & Siminovitch KA (1996) Signaling capacity of the T cell antigen receptor is negatively regulated by the PTP1C tyrosine phosphatase. *J Exp Med.* **184**, 839-852.

117. Joliat MJ, Lang PA, Lyons BL, Burzenski L, Lynes MA, Yi T, Sundberg JP & Shultz LD (2002) Absence of CD5 dramatically reduces progression of pulmonary inflammatory lesions in SHP-1 protein-tyrosine phosphatase-deficient 'viable motheaten' mice. *J Autoimmun.* **18**, 105-117.

118. Van de Velde H & Thielemans K (1996) Native soluble CD5 delivers a costimulatory signal to resting human B lymphocytes. *Cell Immunol.* **172**, 84-91.

119. Mage RG & Pospisil R (2000) CD5 and other superantigens may select and maintain rabbit self-renewing B-lymphocytes and human B-CLL cells. *Curr Top Microbiol Immunol.* **252**, 87-96.

120. Bikah G, Lynd FM, Aruffo AA, Ledbetter JA & Bondada S (1998) A role for CD5 in cognate interactions between T cells and B cells, and identification of a novel ligand for CD5. *Int Immunol.* **10**, 1185-1196.

121. Fruman DA (2004) Phosphoinositide 3-kinase and its targets in B-cell and T-cell signaling. *Curr Opin Immunol.* **16**, 314-320.

122. Gringhuis SI, de Leij LF, Coffer PJ & Vellenga E (1998) Signaling through CD5 activates a pathway involving phosphatidylinositol 3-kinase, Vav, and Rac1 in human mature T lymphocytes. *Mol Cell Biol.* **18**, 1725-1735.

123. Anderson KA & Means AR (2002) Defective signaling in a subpopulation of CD4(+) T cells in the absence of $Ca^{2+}$/calmodulin-dependent protein kinase IV. *Mol Cell Biol.* **22**, 23-29.

124. Raman V, Blaeser F, Ho N, Engle DL, Williams CB & Chatila TA (2001) Requirement for $Ca^{2+}$/calmodulin-dependent kinase type IV/Gr in setting the thymocyte selection threshold. *J Immunol.* **167**, 6270-6278.

125. Gringhuis SI, de Leij LF, Wayman GA, Tokumitsu H & Vellenga E (1997) The $Ca^{2+}$/calmodulin-dependent kinase type IV is involved in the CD5-mediated signaling pathway in human T lymphocytes. *J Biol Chem.* **272**, 31809-31820.

126. Simeoni L, Kliche S, Lindquist J & Schraven B (2004) Adaptors and linkers in T and B cells. *Curr Opin Immunol* . **16**, 304-313.

127. Naramura M, Jang IK, Kole H, Huang F, Haines D & Gu H (2002) c-Cbl and Cbl-b regulate T cell responsiveness by promoting ligand-induced TCR down-modulation. *Nat Immunol.* **3**, 1192-1199.

128. Gamblin SJ & Smerdon SJ (1998) GTPase-activating proteins and their complexes. *Curr Opin Struct Biol.* **8**, 195-201.

129. Dennehy KM, Broszeit R, Ferris WF & Beyers AD (1998) Thymocyte activation induces the association of the proto-oncoprotein c-cbl and ras GTPase-activating protein with CD5. *Eur J Immunol.* **28**, 1617-1625.

130. Sentman CL, Shutter JR, Hockenbery D, Kanagawa O & Korsmeyer SJ (1991) Bcl-2 inhibits multiple forms of apoptosis but not negative selection in thymocytes. *Cell.* **67**, 879-888.

131. Simarro M, Calvo J, Vila JM, Places L, Padilla O, Alberola-Ila J, Vives J & Lozano F (1999) Signaling through CD5 involves acidic sphingomyelinase, protein kinase C-zeta, mitogen-activated protein kinase kinase, and c-Jun NH2-terminal kinase. *J Immunol.* **162**, 5149-5155.

132. Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Res.* **28**, 304-305.

133. Goni FM & Alonso A (2002) Sphingomyelinases: enzymology and membrane activity. *FEBS Lett.* **531**, 38-46.

134. Goni FM & Alonso A (1999) Structure and functional properties of diacylglycerols in membranes. *Prog Lipid Res* . **38**, 1-48.

135. Beyers AD, Spruyt LL & Williams AF (1992) Molecular associations between the T-lymphocyte antigen receptor complex and the surface antigens CD2, CD4, or CD8 and CD5. *Proc Natl Acad Sci USA.* **89**, 2945-2949.

136. Maecker HT, Todd SC & Levy S (1997) The tetraspanin superfamily: molecular facilitators. *FASEB J.* **11**, 428-442.

137. Toyo-oka K, Yashiro-Ohtani Y, Park CS, Tai XG, Miyake K, Hamaoka T & Fujiwara H (1999) Association of a tetraspanin CD9 with CD5 on the T cell surface: role of particular transmembrane domains in the association. *Int Immunol.* **11**, 2043-2052.

138. Davis SJ & van der Merwe PA (1996) The structure and ligand interactions of CD2: implications for T-cell function. *Immunol Today.* **17**, 177-187.

139. Latchman Y & Reiser H (1998) Enhanced murine CD4+ T cell responses induced by the CD2 ligand CD48. *Eur J Immunol.* **28**, 4325-4331.

140. Dustin ML, Golan DE, Zhu DM, Miller JM, Meier W, Davies EA & van der Merwe PA (1997) Low affinity interaction of human or rat T cell adhesion molecule CD2 with its ligand aligns adhering membranes to achieve high physiological affinity. *J Biol Chem.* **272**, 30889-30898.

141. Castro MA, Tavares PA, Almeida MS, Nunes RJ, Wright MD, Mason D, Moreira A & Carmo AM (2002) CD2 physically associates with CD5 in rat T lymphocytes with the involvement of both extracellular and intracellular domains. *Eur J Immunol.* **32**, 1509-1518.

142. Killeen N, Stuart SG & Littman DR (1992) Development and function of T cells in mice with a disrupted CD2 gene. *EMBO J.* **11**, 4329-4336.

143. Teh SJ, Killeen N, Tarakhovsky A, Littman DR & Teh HS (1997) CD2 regulates the positive selection and function of antigen-specific CD4- CD8+ T cells. *Blood.* **89**, 1308-1318.

144. Beyers AD, Barclay AN, Law DA, He Q & Williams AF (1989) Activation of T lymphocytes via monoclonal antibodies against rat cell surface antigens with particular reference to CD2 antigen. *Immunol Rev.* **111**, 59-77.

145. Padilla O, Calvo J, Vila JM, Arman M, Gimferrer I, Places L, Arias MT, Pujana MA, Vives J & Lozano F (2000) Genomic organization of the human CD5 gene. *Immunogenetics.* **51**, 993-1001.

146. Bowen MA, Whitney GS, Neubauer M, Starling GC, Palmer D, Zhang J, Nowak NJ, Shows TB & Aruffo A (1997) Structure and chromosomal location of the human CD6 gene: detection of five human CD6 isoforms. *J Immunol.* **158**, 1149-1156.

147. Sarrias MR, Gronlund J, Padilla O, Madsen J, Holmskov U & Lozano F (2004) The Scavenger Receptor Cysteine-Rich (SRCR) domain: an ancient and highly conserved protein module of the innate immune system. *Crit Rev Immunol.* **24**, 1-37.

148. Hassan NJ, Barclay AN & Brown MH (2004) Frontline: Optimal T cell activation requires the engagement of CD6 and CD166. *Eur J Immunol.* **34**, 930-940.

149. Singer NG, Fox DA, Haqqi TM, Beretta L, Endres JS, Prohaska S, Parnes JR, Bromberg J & Sramkoski RM (2002) CD6: expression during development, apoptosis and selection of human and mouse thymocytes. *Int Immunol.* **14**, 585-597.

150. Osorio LM, Rottenberg M, Jondal M & Chow SC (1998) Simultaneous cross-linking of CD6 and CD28 induces cell proliferation in resting T cells. *Immunology.* **93**, 358-365.

151. Gangemi RM, Swack JA, Gaviria DM & Romain PL (1989) Anti-T12, an anti-CD6 monoclonal antibody, can activate human T lymphocytes. *J Immunol.* **143**, 2439-2447.

152. Bott CM, Doshi JB, Morimoto C, Romain PL & Fox DA (1993) Activation of human T cells through CD6: functional effects of a novel anti-CD6 monoclonal antibody and definition of four epitopes of the CD6 glycoprotein. *Int Immunol.* **5**, 783-792.

153. Osorio LM, De Santiago A, Aguilar-Santelises M, Mellstedt H & Jondal M (1997) CD6 ligation modulates the Bcl-2/Bax ratio and protects chronic lymphocytic leukemia B cells from apoptosis induced by anti-IgM. *Blood.* **89**, 2833-2841.

154. Gimferrer I, Calvo M, Mittelbrunn M, Farnos M, Sarrias MR, Enrich C, Vives J, Sanchez-Madrid F & Lozano F (2004) Relevance of CD6-mediated interactions in T cell activation and proliferation. *J Immunol.* **173**, 2262-2270.

155. Castro MA, Nunes RJ, Oliveira MI, Tavares PA, Simoes C, Parnes JR, Moreira A & Carmo AM (2003) OX52 is the rat homologue of CD6: evidence for an effector function in the regulation of CD5 phosphorylation. *J Leukoc Biol.* **73**, 183-190.

156. Gimferrer I, Farnos M, Calvo M, Mittelbrunn M, Enrich C, Sanchez-Madrid F, Vives J & Lozano F (2003) The accessory molecules CD5 and CD6 associate on the membrane of lymphoid T cells. *J Biol Chem.* **278**, 8564-8571.

157. Pospisil R, Silverman GJ, Marti GE, Aruffo A, Bowen MA & Mage RG (2000) CD5 is A potential selecting ligand for B-cell surface immunoglobulin: a possible role in maintenance and selective expansion of normal and malignant B cells. *Leuk Lymphoma.* **36**, 353-365.

158. Haas KM & Estes DM (2001) The identification and characterization of a ligand for bovine CD5. *J Immunol.* **166**, 3158-3166.

159. Biancone L, Bowen MA, Lim A, Aruffo A, Andres G & Stamenkovic I (1996) Identification of a novel inducible cell-surface ligand of CD5 on activated lymphocytes. *J Exp Med.* **184**, 811-819.

160. Axtell RC, Webb MS, Barnum SR & Raman C (2004) Cutting edge: critical role for CD5 in experimental autoimmune encephalomyelitis: inhibition of engagement reverses disease in mice. *J Immunol* . **173**, 2928-2932.

161. Calvo J, Places L, Espinosa G, Padilla O, Vila JM, Villamor N, Ingelmo M, Gallart T, Vives J, Font J & Lozano F (1999) Identification of a natural soluble form of human CD5. *Tissue Antigens* . **54**, 128-137.

162. Van de Velde H, von Hoegen I, Luo W, Parnes JR & Thielemans K (1991) The B-cell surface protein CD72/Lyb-2 is the ligand for CD5. *Nature.* **351**, 662-665.

163. Luo W, Van d, V, von H, I, Parnes JR & Thielemans K (1992) Ly-1 (CD5), a membrane glycoprotein of mouse T lymphocytes and a subset of B cells, is a natural ligand of the B cell surface protein Lyb-2 (CD72). *J Immunol.* **148**, 1630-1634.

164. Brown MH & Barclay AN (1994) Expression of immunoglobulin and scavenger receptor superfamily domains as chimeric proteins with domains 3 and 4 of CD4 for ligand analysis. *Protein Eng.* **7**, 515-521.

165. Calvo J, Places L, Padilla O, Vila JM, Vives J, Bowen MA & Lozano F (1999) Interaction of recombinant and natural soluble CD5 forms with an alternative cell surface ligand. *Eur J Immunol.* **29**, 2119-2129.

166. Pospisil R, Fitts MG & Mage RG (1996) CD5 is a potential selecting ligand for B cell surface immunoglobulin framework region sequences. *J Exp Med.* **184**, 1279-1284.

167. Pospisil R & Mage RG (1998) CD5 and other superantigens as 'ticklers' of the B-cell receptor. *Immunol Today.* **19**, 106-108.

# Chapter Two

## Renaturation of two extracellular domains from human CD5 expressed in *Escherichia coli*: Optimisation of *in vitro* oxidative folding conditions using a fractional factorial screen


### Abstract

Constructs of the first (*h*CD5d1) and third (*h*CD5d3) scavenger receptor cysteine-rich (SRCR) domains of human CD5 were overexpressed in *Escherichia coli* BL21(DE3) cells. Both constructs expressed at high levels as inclusion bodies. Initial efforts to fold the domains *in vitro* by dialysis or dilution were not successful, due to substantial precipitation. The purification of the reduced unfolded monomers using size-exclusion chromatography prior to *in vitro* folding improved the yield, but still, yields were low. Published protein folding screens were applied but without success. However, these results inspire the design of a new protein folding screen optimised for proteins having disulphide bonds in their native conformation. Reversed phase HPLC was found to be useful as a rapid assay to detect disulphide formation, because a protein containing intra-molecular disulphide bonds is usually more hydrophilic in denaturing conditions than its reduced counterpart. Non-denaturing polyacrylamide gel electrophoresis was also found to be valuable for detecting the change in protein shape due to folding. This new oxidative folding screen was tested with lysozyme and RNase A, and proved successful in finding conditions to fold *h*CD5d1 and *h*CD5d3 with good yields. 1D and 2D NMR showed that both of these constructs have a good dispersion of resonances, which is characteristic of folded proteins.

## Chapter abbreviations and acronyms

**1D** one-dimensional

**2D** two-dimensional

**ε** molar extinction coefficient

**CV** column volume

**EDTA** ethylene diamine tetra acetic acid, di-sodium salt

**DsbA** periplasmic protein thiol:disulphide oxidoreductase from *Escherichia coli*

**DsbC** periplasmic protein thiol:disulphide isomerase from *Escherichia coli*

**GdnHCl** guanidine hydrochloride

**GSH** L-gamma-glutamyl-L-cysteinylglycine (glutathione, reduced form)

**GSSG** glutathione, oxidised form

***h*CD5d1** N-terminal SRCR domain of CD5

***h*CD5d3** C-terminal SRCR domain of CD5

**HPLC** high performance liquid chromatography

**HIC** hydrophobic interaction chromatography

**IDA** iminodiacetic acid

**IPTG** isopropyl-β-D-thiogalactopyranoside

**NMR** nuclear magnetic resonance

**PDB** Protein Data Bank

**PDI** protein disulphide isomerase

**RNase A** bovine pancreatic ribonuclease A

**RPC** reversed phase chromatography

**RT** room temperature

**s** standard deviation

**$s^2$** variance

**SEC** size-exclusion chromatography

## Introduction

The remarkable process of transforming inactive and water insoluble, heat or acid-denatured protein into its biologically active form was first reported by Mortimer Anson and Alfred Mirsky in the early 1930s[1,2]. Nobel laureate crystallographer Max Perutz, then a graduate student, wrote a famous note 'Unboiling an Egg'[3], a title that illustrates how striking this revolutionary discovery was. From the early 1980s, with the development of the technology to over-express heterologous proteins in *Escherichia coli*, the techniques originally used to study the mechanisms of protein renaturation found a new application: the *in vitro* 'activation' of recombinant polypeptide chains that did not express *in vivo* in their biologically active, native, form. Over-expression products that do not attain their native conformation in the cytoplasm often self-aggregate forming insoluble dense particles called inclusion bodies. The inclusion bodies can be isolated by centrifugation of the cell lysate, solubilised using chaotropes (urea, guanidine hydrochloride) or detergents (sarkosyl, Triton X-100, sodium deoxycholate) and folded *in vitro*. The folding is promoted by reducing the concentration of the solubilising agent significantly, so that the denatured polypeptide can rearrange and establish native intramolecular interactions. This process can be accomplished by direct rapid dilution, by dialysis or exchanging buffers using a chromatography matrix (size-exclusion, metal ion affinity). *In vitro* folding is used routinely to prepare proteins for functional and structural studies as well as in large-scale commercial production. Some examples of folded proteins produced in an industrial scale are human insulin[4] for the treatment of diabetes (*Humulin*, Eli Lilly), bovine chymosin[5] for the production of cheese (*Chymax*, CHR Hansen), interleukin-2[6] for the treatment of cancer (*Proleukin*, Chiron) and human interferon-$\alpha$2[7] for leukaemia, hepatitis B and C and AIDS-related Kaposi's sarcoma (*Roferon-A*, Roche; *Intron-A*, Shering-Plough).

### *Inhibition of aggregation during in vitro protein folding*

When the denaturing conditions are withdrawn, most proteins undergo irreversible aggregation as well as folding. Inhibiting protein aggregation is fundamental to achieving a good folding yield. The most straightforward way of doing so is to keep the protein concentration low. However, working at low protein concentrations is often not practical because large volumes of folding buffer are needed and extensive concentration is required afterwards. Several methods have been developed in order to reduce aggregation while keeping the protein concentration relatively high, such strategies include:

i. Using aggregation inhibitors in the folding buffer. Some common examples are arginine[8], low concentrations of chaotropes[9-13], polyethylene glycol (PEG)[14-17], detergents (CHAPS[18], Tween[19], Triton[20] and SDS[21]) and antibodies[22,23].
ii. Folding in the presence of a mixture of protein chaperones and/or isomerases. Examples are immobilised GroEL, DsbA and prolyl isomerase[24,25] and DsbA and DsbC[26].
iii. Mimicking chaperone activity by adding detergent to the unfolded protein and then initiating folding using cyclodextrins[27-30].
iv. Augmenting the total charge of the polypeptide chains by site-directed mutagenesis or, in the case of proteins with cysteines, forming mixed disulphide bonds with polar thiols like glutathione[31].
v. Folding on a solid support to prevent individual polypeptides from interacting with each other. This has been achieved using size-exclusion chromatography[32] and NTA-agarose for poly-histidine tagged proteins[33-37].
vi. Applying high pressure to reverse aggregation[38-40].

The greatest drawback of *in vitro* protein folding is that suitable folding conditions seem to be unpredictably different for every protein. Variables such as optimal temperature, redox potential, pH, and ionic strength and the efficacy of any of the above outlined methodologies to inhibit aggregation, have to be found empirically. This process is usually a very slow and labour intensive task, hence systematic screens that can rapidly assess different conditions are very valuable.

*In vitro protein folding screens*

The first attempt to systematise the search for *in vitro* folding conditions was published by Hofmann and co-workers[41]. They used a matrix of 50 buffers for protein crystallisation to find suitable folding conditions for 9 enzymes using activity assays to quantify the folding yield. Some years latter, Chen and Gouaux[42] designed a protein folding screen based on a fractional factorial design, which was further improved[43] and made commercially available from Hampton Research, under the name of FoldIt. Recently, Klein and co-workers published a modification of the Gouaux screen[44]. Instead of screening for the variables suggested in the original *in vitro* folding protocol, they screened variables they knew from experience to be significant for the folding of the target protein; this flexibility is a general feature of factorial screens and makes them very versatile. Tobbell and co-workers also published a slight modification of the

Gouaux screen[45], their major innovation being that they carried out a first round of screening to detect the important variables, and a second round to optimise the values for such variables.

The Gouaux screen tests the importance for folding of 12 variables (pH, protein concentration, redox potential, ionic strength and the presence of several additives) at two different values for each variable (*e.g.* pH 6.2 and pH 8.2). The screen consists of 16 parallel *in vitro* folding experiments each one with a different combination of values of the 12 variables. If the experimenter has some knowledge of the characteristics and behaviour of the target protein, using a large set of fixed variables that may be beneficial, detrimental or neutral for the folding of their particular protein is not an efficient strategy. Instead, the experimenter could select the particular variables that are expected to promote folding, exclude those that are known to be unfavourable and design a customised smaller parameter set. By doing so, fewer experiments will need to be performed and the probability of achieving good yields will be higher. Recently, Pierce launched a fractional factorial folding screen called Pro-matrix. Pierce made explicit this potential of choosing factors to screen, what they called the 'adjustable matrix'. Pro-matrix screens the effect of only four variables, but two of them are chosen by the experimenter to suit the particular requirements of their target protein. By running 9 experiments, Pro-matrix screens these four variables at three different values, instead of only two, which provides a finer sampling of folding conditions.

*In vitro protein folding with concomitant disulphide bond formation*

Proteins that require disulphide bond formation to fold need an environment that allows thiol oxidation. Oxidation can utilise ambient oxygen in the air[46-48], or more frequently, by using catalytic mixtures of low molecular weight thiols in both their reduced and oxidised form[49]. Such a mixture is known as an oxido-shuffling system. The most popular oxido-shuffling reagents are oxidised and reduced glutathione (GSSG/GSH). Other pairs are cysteine/cystine and 2-mercaptoethanol/2-hydroxyethyldisulphide. Little has been reported using other redox couples. However, some more efficient reagents have been developed recently, including 1,2-bis(mercaptoacetamido)cyclohexane (Vectrase-P)[50] and aromatic thiols such as 4-mercaptobenzeneacetate[51]. These new molecules are able to increase both the rate of the folding reaction and the final yield of folded protein. The superior performance appears to be the result of having similar thiol pKa and redox potential to that of the eukaryotic protein disulphide isomerase (PDI)[50,51],

and in the case of Vectrase-P also to the presence of two thiol groups in the same molecule, which is another characteristic of PDI [52].

All oxido-shuffling reagents undergo an extensive thiol/disulphide exchange process with the cysteines of the protein. First a mixed disulphide bond is formed between the oxido-shuffling reagent and one of the free cysteines in the protein. Then another of the protein's cysteines displaces the oxido-shuffling reagent to form an intramolecular disulphide bond[53,54] (Figure II.1). The thiol/disulphide exchange is a classical bimolecular nucleophilic substitution ($S_N2$), where the active species is the thiolate ion ($S^-$)[54,55]. Therefore the reaction rate is favoured at basic pH values, above the pKa of the sulphydryl group. The involved thiols have to be good nucleophiles to attack other thiols, but they also have to be good leaving groups so that they can, in turn, be displaced by another thiol. For this reason, the optimal pH for the reaction is close to the pKa of the thiols involved[52,56]. Both the reduced and the oxidised forms are needed because the formation of native disulphide bonds requires extensive rearrangement, so that a continuous cycle of reduction and oxidation takes place until the native disulphide bonds are formed[53]. This situation has been observed not only with oxido-shuffling reagents, but also with the oxido-reductase chaperone machinery in bacteria[57] and eukaryotes[58]. In bacteria, the oxidation of free cysteines is performed by DsbA, while the isomerisation of non-native disulphide bonds is facilitated by DsbC. Both *in vivo* and *in vitro*, the presence of DsbC but not DsbA enhances the efficiency of folding[26].

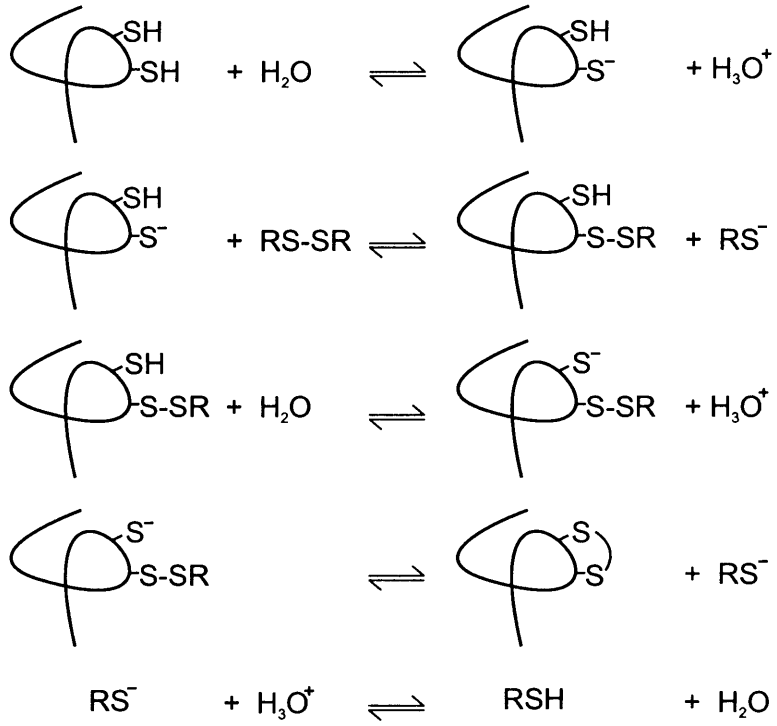


**Figure II.1** Representation of the reactions involved in the formation of protein disulphide bonds in the presence of an oxido-shuffling pair (RSH/ RSSR). The reaction rate is favoured at basic pH values, above the pKa of the involved thiol groups.

*Choice of variables to screen for protein folding with disulphide bond formation*

Important variables that any folding screen should include are protein concentration, temperature and the presence of arginine and guanidine hydrochloride (GdnHCl), which have been shown to be by far the most successful folding additives. Given the industrial and basic research importance of *in vitro* protein folding, the literature describing *in vitro* folding methodologies is vast. Several reviews have been published gathering tens of different conditions that have been successful for folding proteins with concomitant disulphide bond formation[59-63]. All this information is extremely valuable in choosing variables to screen and the values for those variables. By reviewing the literature the following guidelines can be established:

i. The folding buffer pH has to be basic, usually between pH 8.0 and 10.0.
ii. Substrate protein concentration ranges mainly between 10 μg/ml and 100 μg/ml.
iii. The ratio of reduced to oxidised oxido-shuffling thiol reagent is between 10:1 and 1:1 with a reduced thiol concentration between 1 mM and 10 mM.
iv. Widely used low molecular weight folding enhancers are:
    a. Arginine, at 0.4 M to 1.0 M concentration
    b. Urea or GdnHCl, at 1 M to 2 M concentration
    c. Polyethylene glycol (PEG)
    d. Glycerol
    e. Detergents (*e.g.* CHAPS and Triton X-100)
    f. There are no general trends in folding temperature or time. These parameters range from 4°C to 37°C and from 1 hr to 150 hrs.

*General techniques to characterise protein folding*

In the quest for a folding screen that would be widely applicable, a general method to assess the amount of native folded protein is a fundamental requirement. The ideal method to measure folding yields has to be straightforward, fast, preferably quantitative, and as insensitive to methodological variations as possible. These requirements rule out the traditional methods to quantify *in vitro* folding yield because they rely on substrate-specific methods such as catalytic activity[41,43,45] or conformation-dependent antibody or ligand binding[43,64]. Heiring and Muller[65] demonstrated the use of limited proteolysis followed by SDS PAGE and silver staining as a universal method to distinguish between folded and unfolded protein, however, this method is not quantitative and requires a great deal of sample manipulation, which increases the possibility of experimental error.

Alternatively, spectroscopic techniques such as intrinsic fluorescence[66] could be used if readily available. In other circumstances chromatographic techniques are the method of choice. Armstrong and co-workers[43] introduced analytical size-exclusion (SEC) chromatography to quantify folding yield. However, SEC is very restricted in terms of the sample volume size, as the elution peaks start to broaden with sample volumes larger that 2% the volume of the column. Therefore, the protein concentration of the loaded sample has to be fairly high in order to detect the eluted protein. As this is often not the case after a protein folding screen, the volume limitation is a major drawback to this method. Another contribution was made recently by Scheich and co-workers[66], who quantified folding yields using hydrophobic interaction chromatography (HIC), thereby making use of the fact that folded proteins are less hydrophobic than their unfolded counterparts. In HIC there is no restriction on the sample volume, if the protein concentration is very low, it is possible to load a greater volume in order to be able to detect it. This enrichment aspect of HIC is an enormous advantage over SEC.

*Screening variables using factorial designs*

Knowledge of experimental design and statistical analysis is very valuable in the process of solving problems and interpreting the results of experiments involving many variables. Among the experimental problems that can be approached using statistics, two are particularly relevant to the field of recombinant protein expression and purification:

a. The selection of conditions to maximise yields, *e.g.* protein solubility or protein expression levels, by systematic identification of those variables that are important for the outcome of a process.
b. The evaluation of the magnitude and sources of variation in measurements with the aim to determine whether the data obtained truly provide information about the system or simply reflect experimental or measurement errors.

In screening for suitable protein folding conditions, the experimenter needs to assess the effect of many variables, individually or jointly, on the final folding yield. One of the ways that experimental design can address this kind of problem is with a factorial experiment. A factorial strategy consists of systematically testing the influence on the experimental outcome of every possible combination of the variables to be studied, and by doing so, determine which variables have an statistically significant positive or negative influence. Published examples of systems that have been optimised using some form of factorial screen include: protein crystallisation[67,68], recombinant and endogenous

protein expression[69-72], protein chromatography[73], cell growth[74,75], and cDNA microarray signal-to-noise ratio[76].

As discussed above, there are many variables (termed *factors*[1] in the experimental design field), that may or may not improve folding yields. These factors include protein concentration, temperature, time, pH, nature, ratio and concentration of oxido-shuffling agents, and the presence of folding enhancers. Despite the justifiable desire to test all these variables at once, when conducting a factorial screen it is important to keep the number of factors low because the number of combinations or *runs* needed to screen these factors increases rapidly, from only 4 runs for 2 factors to 256 runs for 8 factors. Once the factors have been chosen, two or more values (*levels*) for each factor have to be selected. The simplest factorial experiments use only two levels, usually represented by the signs + and – for the extreme high and low levels of each factor respectively. In this project, six factors at two levels, as shown in Table II.1, were chosen for the factorial folding screen.

*Illustration of the factorial design methodology*

Table II.2 shows a set of hypothetical results obtained for a full-factorial folding screen of six factors at two levels. The *response*, the outcome of the experiment that is to be measured, is the area under the peak corresponding to the oxidised, folded, protein in the imaginary chromatogram. The units of the response in this case are arbitrary units of area (AU). Each one of the 64 factor-level combinations is called a *run* or *treatment*. It is an important requisite that runs are assayed randomly, in order to average out potential unaccounted influences on the experiments.

*Calculation of main effects*

The effect of the presence of, for example, arginine (factor A) can be calculated from the difference in the response for run 1 and 2. The only difference between these two runs is that in run 2 the arginine concentration is high (A+; 0.6 M) and in run 1 its concentration is low (A-; 0.0 M); all of the other factors remain constant. In a similar manner, one could also calculate the effect of arginine in different backgrounds by subtracting the response of run 3 from 4, run 5 from 6, run 7 from 8 and so on. The *main effect* of arginine is the average for all 32 individual effects. A simpler way of calculating the

[1] Italics will be used to introduce experimental design terminology for the first time

same main effect is to subtract the average yield of all runs in which arginine was at its higher level from the average yield of all runs in which arginine was at its lower level. Taking n as the number of runs:

$$\begin{aligned}\text{Arginine effect} &= \tfrac{2}{n}\sum yield(A+) - \tfrac{2}{n}\sum yield(A-) \\ &= \tfrac{2}{64}(3595) - \tfrac{2}{64}(3247) \\ &= 10.9 \text{ AU}\end{aligned}$$

In this instance, interpretation is that the presence of arginine, at a concentration of 0.6 M, has an average positive effect of 10.9 AU on the yield of folded protein. The other main effects can be derived in a similar manner, and their values appear in Table II.3. In this example, arginine, and glycerol have a positive effect on the folded protein yield and GdnHCl, high temperature, high pH and high protein concentration have a negative effect.

*Calculation of interaction effects*

If instead of a factorial design, the experimenter used a 'change-one-factor-at-a-time' strategy, it would not be possible to tell whether the observed effect of a particular factor depends at all on the fixed levels for all the other factors. In this example, the averaged main effect of temperature (D) is slightly negative -1.9, meaning that folding at room temperature (RT) instead of 4°C would reduce the yield by 1.9 AU. However, the average effect of temperature when the protein concentration is low is 6.1 AU, and the average effect of temperature when the protein concentration is high is -10.0 AU. At high protein concentration, the effect of temperature is to dramatically decrease the yield, but at low protein concentration, the yield actually increases if the experiment is performed at room temperature. The effects of temperature and protein concentration depend on each other; they are said to be *interacting*. Half the difference between the average temperature effect with high protein concentration and the average temperature effect with low protein concentration is called the 'temperature by protein concentration *interaction*' (denoted DF using the symbols from Table II.1). In this case DF has a value of -8.1 AU. As can be seen in Table II.3, there are no other important two-factor interaction effects.

**Table II.1 Factors and levels for the *in vitro* folding screen**

| | Factor | Level | |
|---|---|---|---|
| | | + | - |
| A | [Arginine] | 0.6 M | 0.0 M |
| B | [Guanidine HCl] | 0.5 M | 0.0 M |
| C | [Glycerol] (w/v) | 20% | 0% |
| D | Temperature | Room temp | 4°C |
| E | pH | 9.0 | 8.0 |
| F | Protein concentration | 0.2 mg/ml | 0.1 mg/ml |

**Table II.2 Design matrix for a $2^6$ full-factorial *in vitro* folding screen and hypothetical experimental results**

| | Factor* | | | | | | | Area** | | | | | Factor | | | | | | | Area | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | A | B | C | D | E | F | Order‡ | R1 | R2 | Average | $s^2$ | Run | A | B | C | D | E | F | Order‡ | R1 | R2 | Average | $s^2$ |
| 1 | - | - | - | - | - | - | 14 | 108 | 112 | 110 | 8 | 33 | - | - | - | - | - | + | 19 | 100 | 104 | 102 | 8 |
| 2 | + | - | - | - | - | - | 61 | 127 | 123 | 125 | 8 | 34 | + | - | - | - | - | + | 6 | 111 | 105 | 108 | 18 |
| 3 | - | + | - | - | - | - | 60 | 112 | 112 | 112 | 0 | 35 | - | + | - | - | - | + | 10 | 90 | 94 | 92 | 8 |
| 4 | + | + | - | - | - | - | 38 | 120 | 124 | 122 | 8 | 36 | + | + | - | - | - | + | 1 | 108 | 102 | 105 | 18 |
| 5 | - | - | + | - | - | - | 33 | 121 | 113 | 117 | 32 | 37 | - | - | + | - | - | + | 43 | 93 | 85 | 89 | 32 |
| 6 | + | - | + | - | - | - | 55 | 121 | 127 | 124 | 18 | 38 | + | - | + | - | - | + | 27 | 96 | 104 | 100 | 32 |
| 7 | - | + | + | - | - | - | 13 | 116 | 108 | 112 | 32 | 39 | - | + | + | - | - | + | 59 | 86 | 96 | 91 | 50 |
| 8 | + | + | + | - | - | - | 8 | 125 | 129 | 127 | 8 | 40 | + | + | + | - | - | + | 62 | 106 | 100 | 103 | 18 |
| 9 | - | - | - | + | - | - | 16 | 110 | 120 | 115 | 50 | 41 | - | - | - | + | - | + | 31 | 88 | 82 | 85 | 18 |
| 10 | + | - | - | + | - | - | 50 | 135 | 125 | 130 | 50 | 42 | + | - | - | + | - | + | 11 | 87 | 89 | 88 | 2 |
| 11 | - | + | - | + | - | - | 56 | 112 | 124 | 118 | 72 | 43 | - | + | - | + | - | + | 28 | 73 | 85 | 79 | 72 |
| 12 | + | + | - | + | - | - | 48 | 141 | 125 | 133 | 128 | 44 | + | + | - | + | - | + | 58 | 91 | 87 | 89 | 8 |
| 13 | - | - | + | + | - | - | 51 | 117 | 121 | 119 | 8 | 45 | - | - | + | + | - | + | 35 | 89 | 85 | 87 | 8 |
| 14 | + | - | + | + | - | - | 29 | 137 | 127 | 132 | 50 | 46 | + | - | + | + | - | + | 20 | 93 | 85 | 89 | 32 |
| 15 | - | + | + | + | - | - | 37 | 114 | 128 | 121 | 98 | 47 | - | + | + | + | - | + | 17 | 82 | 88 | 85 | 18 |
| 16 | + | + | + | + | - | - | 30 | 128 | 130 | 129 | 2 | 48 | + | + | + | + | - | + | 42 | 90 | 86 | 88 | 8 |
| 17 | - | - | - | - | + | - | 45 | 110 | 116 | 113 | 18 | 49 | - | - | - | - | + | + | 4 | 93 | 87 | 90 | 18 |
| 18 | + | - | - | - | + | - | 40 | 124 | 120 | 122 | 8 | 50 | + | - | - | - | + | + | 7 | 106 | 102 | 104 | 8 |
| 19 | - | + | - | - | + | - | 3 | 110 | 102 | 106 | 32 | 51 | - | + | - | - | + | + | 12 | 86 | 98 | 92 | 72 |
| 20 | + | + | - | - | + | - | 47 | 121 | 125 | 123 | 8 | 52 | + | + | - | - | + | + | 32 | 100 | 104 | 102 | 8 |
| 21 | - | - | + | - | + | - | 24 | 111 | 117 | 114 | 18 | 53 | - | - | + | - | + | + | 25 | 99 | 91 | 95 | 32 |
| 22 | + | - | + | - | + | - | 9 | 120 | 124 | 122 | 8 | 54 | + | - | + | - | + | + | 23 | 107 | 101 | 104 | 18 |
| 23 | - | + | + | - | + | - | 21 | 103 | 111 | 107 | 32 | 55 | - | + | + | - | + | + | 53 | 99 | 93 | 96 | 18 |
| 24 | + | + | + | - | + | - | 18 | 126 | 120 | 123 | 18 | 56 | + | + | + | - | + | + | 22 | 105 | 95 | 100 | 50 |
| 25 | - | - | - | + | + | - | 5 | 117 | 115 | 116 | 2 | 57 | - | - | - | + | + | + | 49 | 80 | 90 | 85 | 50 |
| 26 | + | - | - | + | + | - | 52 | 124 | 130 | 127 | 18 | 58 | + | - | - | + | + | + | 26 | 95 | 89 | 92 | 18 |
| 27 | - | + | - | + | + | - | 57 | 117 | 111 | 114 | 18 | 59 | - | + | - | + | + | + | 34 | 86 | 82 | 84 | 8 |
| 28 | + | + | - | + | + | - | 36 | 131 | 127 | 129 | 8 | 60 | + | + | - | + | + | + | 46 | 98 | 90 | 94 | 32 |
| 29 | - | - | + | + | + | - | 2 | 124 | 112 | 118 | 72 | 61 | - | - | + | + | + | + | 39 | 80 | 84 | 82 | 8 |
| 30 | + | - | + | + | + | - | 44 | 124 | 132 | 128 | 32 | 62 | + | - | + | + | + | + | 63 | 96 | 90 | 93 | 18 |
| 31 | - | + | + | + | + | - | 15 | 120 | 116 | 118 | 8 | 63 | - | + | + | + | + | + | 54 | 80 | 86 | 83 | 18 |
| 32 | + | + | + | + | + | - | 64 | 127 | 133 | 130 | 18 | 64 | + | + | + | + | + | + | 41 | 105 | 115 | 110 | 50 |

* Factors as given in Table II.1

** Area under the elution peak of the UV absoption curve used as a measure of protein quantity. Given in arbitrary units.

‡ Randomised order in which to perform the experiments

R1 and R2 are repeats of the experiment; $s^2$ is the sample variance

A quick method to calculate two and higher-factor effects is to use the table of signs that is shown in Table II.2. To calculate the effect of DF, for example, one multiplies the signs in column **D** by the signs in column **F** by the output response value for each run. These results are added and finally divided by half the number of *runs*.

$$DF_{effect} = \frac{+110+125+112+122+117+124+112+127-115-130-118-...}{32} = \frac{-258}{32} = -8.1$$

In total, a six-factor factorial experiment has 64 effects: the average response, 6 main effects, and 57 interaction effects: 15 two-factor, 20 three-factor, 15 four-factor, 6 five-factor and 1 six-factor. However, as can be seen in Table II.3, *high-order* interactions, *i.e.* those involving three or more factors, are seldom significant. It is a fundamental empirical principle that lower order effects are more likely to be important than higher order effects, a feature of these systems known as hierarchical ordering[77]. In this manner, main effects tend to be more significant than two-factor interactions, which in turn tend to be more important than three-factor interactions, and so on.

*Measuring experimental error*

The standard deviation ($s$) for the effects in a factorial screen can be estimated in different ways. One of them is to perform two or more complete independent experiments containing the full set of runs, each one of these complete experiments is known as a *replicate*. If two replicates are performed, the variance ($s^2$) for each run ($i$) is the square of the difference (**d**) between the duplicate observations divided by two:

$$s_i^2 = \mathbf{d}_i^2 / 2$$

The variance of each effect is equal to four times the pooled run variance (the sum of the individual run variances divided by the number of runs in each replicate, n) divided by the total number of runs performed (N= n x number of replicates):

$$s^2_{effects} = \frac{4}{N} \frac{\sum_{i=1}^{n} s_i^2}{n}$$

**Table II.3 Calculated effects for the $2^6$ full-factorial *in vitro* folding screen**

| | Effect (AU) | | Effect (AU) |
|---|---|---|---|
| 1 [Arginine] (A) | 10.9 ± 0.9 | 33 BCE | 0.1 ± 0.9 |
| 2 [GdnHCl] (B) | -0.3 ± 0.9 | 34 BCF | 1.3 ± 0.9 |
| 3 [Glycerol] (C) | 0.9 ± 0.9 | 35 BDE | 0.9 ± 0.9 |
| 4 Temperature (D) | -1.9 ± 0.9 | 36 BDF | 0.0 ± 0.9 |
| 5 pH (E) | -0.3 ± 0.9 | 37 DEF | 1.4 ± 0.9 |
| 6 [Protein] (F) | -27.2 ± 0.9 | 38 CDE | -0.4 ± 0.9 |
| 7 AB | 1.4 ± 0.9 | 39 CDF | 1.2 ± 0.9 |
| 8 AC | -0.4 ± 0.9 | 40 CEF | 1.3 ± 0.9 |
| 9 AD | -0.1 ± 0.9 | 41 DEF | 1.2 ± 0.9 |
| 10 AE | 1.0 ± 0.9 | 42 ABCD | -0.2 ± 0.9 |
| 11 AF | -1.4 ± 0.9 | 43 ABCE | 0.4 ± 0.9 |
| 12 BC | 0.9 ± 0.9 | 44 ABCF | -0.2 ± 0.9 |
| 13 BD | 1.4 ± 0.9 | 45 ABDE | 0.8 ± 0.9 |
| 14 BE | 0.6 ± 0.9 | 46 ABDF | 1.4 ± 0.9 |
| 15 BF | 0.3 ± 0.9 | 47 ABEF | -0.9 ± 0.9 |
| 16 CD | 1.2 ± 0.9 | 48 ACDE | 1.5 ± 0.9 |
| 17 CE | 0.9 ± 0.9 | 49 ACDF | 0.9 ± 0.9 |
| 18 CF | -0.7 ± 0.9 | 50 ACEF | 0.3 ± 0.9 |
| 19 DE | 1.3 ± 0.9 | 51 ADEF | 1.5 ± 0.9 |
| 20 DF | -8.1 ± 0.9 | 52 BCDE | 1.4 ± 0.9 |
| 21 EF | 1.9 ± 0.9 | 53 BCDF | 0.3 ± 0.9 |
| 22 ABC | 0.2 ± 0.9 | 54 BCEF | -0.5 ± 0.9 |
| 23 ABD | 0.3 ± 0.9 | 55 BDEF | 0.5 ± 0.9 |
| 24 ABE | 0.6 ± 0.9 | 56 CDEF | -1.2 ± 0.9 |
| 25 ABF | 0.2 ± 0.9 | 57 ABCDE | 0.9 ± 0.9 |
| 26 ACD | 0.4 ± 0.9 | 58 ABCDF | 0.9 ± 0.9 |
| 27 ACE | 0.6 ± 0.9 | 59 ABCEF | 1.1 ± 0.9 |
| 28 ACF | 0.8 ± 0.9 | 60 ABDEF | 0.9 ± 0.9 |
| 29 ADE | 1.1 ± 0.9 | 61 ACDEF | 1.3 ± 0.9 |
| 30 ADF | -0.3 ± 0.9 | 62 BCDEF | 0.9 ± 0.9 |
| 31 AEF | 1.0 ± 0.9 | 63 ABCDEF | -0.1 ± 0.9 |
| 32 BCD | 0.0 ± 0.9 | 64 Average | 106.90 |

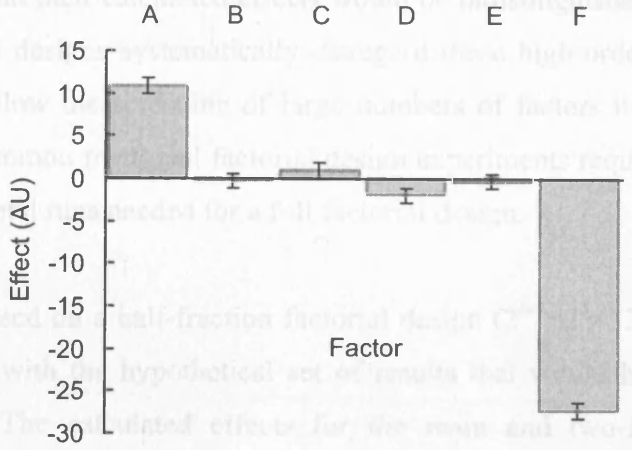


**Figure II.2** Plot of main effects for the $2^6$ full-factorial folding screen shown in Table II.3

For the hypothetical folding screen,

$$s^2_{effects} = \frac{4}{128}\frac{1666}{64}$$

$$s^2_{effects} = 0.81$$

Finally, the estimated standard error in this hypothetical experiment is the square root of the effect variance, 0.9 in this case. Taking this value into account, it is evident that only arginine and protein concentration have a significant effect on the response (Table II.3 and Figure II.2). As there is no evidence that arginine has a significant interaction with any other factor, the experimenter can therefore reliable conclude that the effect of the presence of arginine at its high level is to increase the folding yield and that this effect in not dependent on the other factors. The main effects of both elevated temperature and higher protein concentration are negative. However these two parameters should be considered jointly, because the two-factor interaction effect DF has a significant value. It would be misleading to conclude that the effect of temperature is to decrease the final yield, because, as discussed previously, the effect of temperature is dependent on the protein concentration level.

*Fractional factorial designs*

The greatest drawback of full factorial experiments is that as the number of factors increase, the number of runs rapidly becomes very large. A 8-factor experiment, for example, would require 256 runs. Of the 256 calculated effects 219 would be of interactions involving three or more factors. As the hierarchical ordering principle indicates, these interactions would probably have little or no significance on the outcome of the experiment and their calculated effects would be indistinguishable from the noise. 'Fractional' factorial designs systematically disregard these high-order interactions, and by doing so they allow the screening of large numbers of factors in a relatively small number of runs. Common fractional factorial design experiments require only 1/2, 1/4 or 1/8 of the experimental runs needed for a full factorial design.

A folding screen based on a half-fraction factorial design ($2^{6-1}=2^5=32$ runs) is shown in Table II.4 together with the hypothetical set of results that would have been obtained using this screen. The calculated effects for the main and two-factor interactions, obtained following the same methodology used for the full-factorial design are shown in Table II.5. The analysis of this reduced dataset also highlights that arginine and the

combination of temperature and protein concentration are the important factors. Nevertheless, there is a loss of information. This deficit becomes evident when trying to calculate three or higher order interactions by multiplying the signs in the respective columns.

The sequence of resultant signs obtained for the ABCE interaction by multiplying the signs in columns **A**, **B**, **C** and **E** for each run is:

**ABCE**= + – – + –+ + – + – – + – + + – –+ + – + – – + – + + – + – – +

This result is identical to the sequence of signs for the DF interaction,

**DF**= + – – + –+ + – + – – + – + + – –+ + – + – – + – + + – + – – +

It is said that the ABCE effect and DF effect are *aliases* of each other. The value calculated for the effect of DF in Table II.5 is in fact the sum of the DF and the ABCE effects seen in the full factorial case (Table II.3).

DF + ABCE = -7.6 ± 1.1

Strictly speaking it is not possible to know whether the negative effect calculated in the fractional factorial design is due to the combined effect of temperature and protein concentration, or to the combined effect of the rest of the factors under investigation, or to both. However, the hierarchical order principle states that two-factor interactions are probably more significant than four-factor interactions. Thus, it is reasonable to assume that the calculated effect is mainly due to the two factor interaction of temperature and pH. Of course, this assumption might not be true in the general case, and this is the risk that has to be taken in adopting the fractional factorial design to reduce the experimental effort, time and costs.

**Table II.4 Design matrix for a $2^{6-1}$ half-fraction factorial *in vitro* folding screen and hypothetical experimental results**

| | | Factor | | | | | | | Area | | | | | Factor | | | | | | | Area | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | # | A | B | C | D | E | F | Order‡ | R1 | R2 | Average | $s^2$ | Run | # | A | B | C | D | E | F | Order‡ | R1 | R2 | Average | $s^2$ |
| 1 | 1 | - | - | - | - | - | - | 30 | 108 | 112 | 110 | 8 | 17 | 34 | - | - | - | - | + | + | 10 | 93 | 87 | 90 | 18 |
| 2 | 4 | + | - | - | - | - | + | 8 | 111 | 105 | 108 | 18 | 18 | 35 | + | - | - | - | + | - | 18 | 124 | 120 | 122 | 8 |
| 3 | 6 | - | + | - | - | - | + | 7 | 90 | 94 | 92 | 8 | 19 | 37 | - | + | - | - | + | - | 32 | 110 | 102 | 106 | 32 |
| 4 | 7 | + | + | - | - | - | - | 5 | 120 | 124 | 122 | 8 | 20 | 40 | + | + | - | - | + | + | 19 | 100 | 104 | 102 | 8 |
| 5 | 10 | - | - | + | - | - | + | 25 | 93 | 85 | 89 | 32 | 21 | 41 | - | - | + | - | + | - | 31 | 111 | 117 | 114 | 18 |
| 6 | 11 | + | - | + | - | - | - | 1 | 121 | 127 | 124 | 18 | 22 | 44 | + | - | + | - | + | + | 2 | 107 | 101 | 104 | 18 |
| 7 | 13 | - | + | + | - | - | - | 11 | 116 | 108 | 112 | 32 | 23 | 46 | - | + | + | - | + | + | 17 | 99 | 93 | 96 | 18 |
| 8 | 16 | + | + | + | - | - | + | 15 | 106 | 100 | 103 | 18 | 24 | 47 | + | + | + | - | + | - | 22 | 126 | 120 | 123 | 18 |
| 9 | 18 | - | - | - | + | - | + | 14 | 88 | 82 | 85 | 18 | 25 | 49 | - | - | - | + | + | - | 6 | 117 | 115 | 116 | 2 |
| 10 | 19 | + | - | - | + | - | - | 23 | 135 | 125 | 130 | 50 | 26 | 52 | + | - | - | + | + | + | 16 | 95 | 89 | 92 | 18 |
| 11 | 21 | - | + | - | + | - | - | 9 | 112 | 124 | 118 | 72 | 27 | 54 | - | + | - | + | + | + | 26 | 86 | 82 | 84 | 8 |
| 12 | 24 | + | + | - | + | - | + | 29 | 91 | 87 | 89 | 8 | 28 | 55 | + | + | - | + | + | - | 12 | 131 | 127 | 129 | 8 |
| 13 | 25 | - | - | + | + | - | - | 3 | 117 | 121 | 119 | 8 | 29 | 58 | - | - | + | + | + | + | 13 | 80 | 84 | 82 | 8 |
| 14 | 28 | + | - | + | + | - | + | 24 | 93 | 85 | 89 | 32 | 30 | 59 | + | - | + | + | + | - | 4 | 124 | 132 | 128 | 32 |
| 15 | 30 | - | + | + | + | - | + | 21 | 82 | 88 | 85 | 18 | 31 | 61 | - | + | + | + | + | - | 20 | 120 | 116 | 118 | 8 |
| 16 | 31 | + | + | + | + | - | - | 28 | 128 | 130 | 129 | 2 | 32 | 64 | + | + | + | + | + | + | 27 | 105 | 115 | 110 | 50 |

# Run number in the full-factorial design (Table II.2)
* Factors as given in Table II.1
** Area under the elution peak of the UV absoption curve used as a measure of protein quantity. Given in arbitrary units
‡ Randomised order in which to perform the experiments
R1 and R2 are repeats of the experiment; $s^2$ is the sample variance

**Table II.5 Calculated effects for the $2^{6-1}$ half-fraction factorial *in vitro* folding screen**

| | Effect (AU) | | | Effect (AU) | |
|---|---|---|---|---|---|
| 1 [Arginine] (A) | + BCDEF | 11.8 ± 1.1 | 17 CE | + ABDF | 2.4 ± 1.1 |
| 2 [GdnHCl] (B) | + ACDEF | 1.0 ± 1.1 | 18 CF | + ABDE | 0.1 ± 1.1 |
| 3 [Glycerol] (C) | + ABDEF | 1.9 ± 1.1 | 19 DE | + ABCF | 1.1 ± 1.1 |
| 4 Temperature (D) | + ABCEF | -0.9 ± 1.1 | 20 DF | + ABCE | -7.6 ± 1.1 |
| 5 pH (E) | + ABCDF | 0.8 ± 1.1 | 21 EF | + ABCD | 1.8 ± 1.1 |
| 6 [Protein] (F) | + ABCDE | -26.3 ± 1.1 | 22 ABC | + DEF | 1.4 ± 1.1 |
| 7 AB | + CDEF | 0.3 ± 1.1 | 23 ABD | + CEF | 1.6 ± 1.1 |
| 8 AC | + BDEF | 0.1 ± 1.1 | 24 ABE | + CDF | 1.8 ± 1.1 |
| 9 AD | + BCEF | -0.6 ± 1.1 | 25 ABF | + CDE | -0.3 ± 1.1 |
| 10 AE | + BCDF | 1.3 ± 1.1 | 26 ACD | + BEF | 1.8 ± 1.1 |
| 11 AF | + BCDE | 0.0 ± 1.1 | 27 ACE | + BDF | 0.6 ± 1.1 |
| 12 BC | + ADEF | 2.4 ± 1.1 | 28 ACF | + BDE | 1.6 ± 1.1 |
| 13 BD | + ACEF | 1.6 ± 1.1 | 29 ADE | + BCF | 2.4 ± 1.1 |
| 14 BE | + ACDF | 1.5 ± 1.1 | 30 ADF | + BCE | -0.1 ± 1.1 |
| 15 BF | + ACDE | 1.8 ± 1.1 | 31 AEF | + BCD | 1.0 ± 1.1 |
| 16 CD | + ABEF | 0.3 ± 1.1 | 32 Average | + ABCDEF | 106.9 |

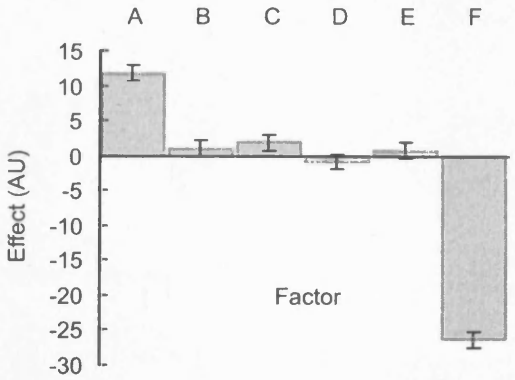


**Figure II.3** Plot of main effects for the $2^{6-1}$ half-fraction factorial folding screen shown in Table II.5

In a full factorial design, each column of signs and each of the 64 effects is unique. In a fractional factorial design all columns have one or more aliases. As can be seen in Table II.5, in a half-fraction design each effect has an alias. Main and two factor interactions are only aliased with three or higher factor interactions, and are therefore said to be *clear*. As fewer and fewer runs are performed to screen for the same number of variables, the aliasing increases. A quarter-fraction design requires 16 runs to screen for 6 factors. In this design each effect has three aliases. The main effects are still clear, but the two-factor interactions are aliased with other two-factor interactions, and they are not clear. An eighth-fraction design requires only 8 runs to screen the same 6 factors ($2^{6-3}= 2^3= 8$ runs). At this point, the design is heavily aliased. Each effect has eight aliases and the not all of the main effects are clear, as some of them are aliased with the two-factor interactions. Table II.7 displays the calculated effects for the eighth fraction factorial design of Table II.6. Arginine and protein concentration are still the variables that show an important effect. However, at this degree of aliasing, it is not possible to identify the interaction between the temperature and the protein concentration. As long as this limitation is kept in mind, an eighth-fraction design is a very attractive option for a folding screen. It provides sufficient information to shed light on the variables that have an important effect on the yield, and it is a good starting point from which one could, if considered necessary, optimise the conditions or study them in greater depth.

**Table II.6 Design matrix for a $2^{6-3}$ eighth-fraction factorial *in vitro* folding screen and hypothetical results**

| | | Factor* | | | | | | | Area | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Run | # | A | B | C | D | E | F | Order‡ | R1 | R2 | Average | $s^2$ |
| 1 | 57 | - | - | - | + | + | + | 5 | 80 | 90 | 85 | 50 |
| 2 | 34 | + | - | - | - | - | + | 3 | 111 | 105 | 108 | 18 |
| 3 | 19 | - | + | - | - | + | - | 8 | 110 | 102 | 106 | 32 |
| 4 | 12 | + | + | - | + | - | - | 4 | 141 | 125 | 133 | 128 |
| 5 | 13 | - | - | + | + | - | - | 7 | 117 | 121 | 119 | 8 |
| 6 | 22 | + | - | + | - | + | - | 6 | 120 | 124 | 122 | 8 |
| 7 | 39 | - | + | + | - | - | + | 1 | 86 | 96 | 91 | 50 |
| 8 | 64 | + | + | + | + | + | + | 2 | 105 | 115 | 110 | 50 |

# Run number in the full-factorial design (Table II.2)
* Factors as given in Table II.1
** Area under the elution peak of the UV absoption curve used as a measure of protein quantity. Given in arbitrary units.
‡ Randomised order in which to perform the experiments
R1 and R2 are repeats of the experiment; $\hat{s}$ is the sample variance

**Table II.7 Calculated effects for the $2^{6-3}$ eighth-fraction factorial *in vitro* folding screen**

| | | Effect (AU) |
|---|---|---|
| 1 [Arginine] (A) | | 9.0 ± 3.3 |
| 2 [GdnHCl] (B) | | 0.8 ± 3.3 |
| 3 [Glycerol] (C) | | 1.3 ± 3.3 |
| 4 Temperature (D) | + AB | 2.5 ± 3.3 |
| 5 pH (E) | + AC | -3.5 ± 3.3 |
| 6 [Protein] (F) | + BC | -10.8 ± 3.3 |

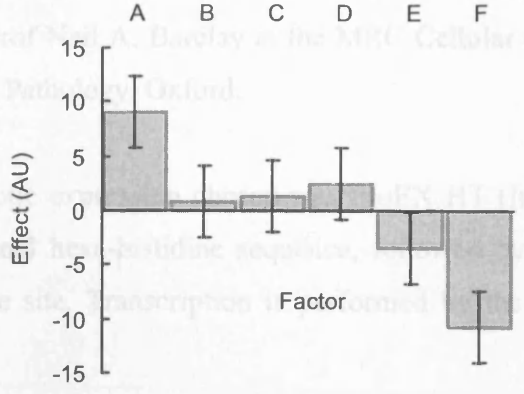


**Figure II.3.** Plot of main effects for the $2^{6-3}$ eighth-fraction factorial folding screen shown in Table II.6

## Experimental procedures[2]

*E. coli expression plasmid for hCD5d1*

The expression plasmid for *h*CD5d1 was constructed by Dr Mark McAlister before I joined the Driscoll group. The construct (Figure II.5) comprises residues 25 to 134 of *h*CD5 inserted between the NdeI and XhoI restriction sites of pET15b (Novagen). The vector incorporates an N-terminal hexa-histidine sequence to allow for immobilised metal affinity purification of the expression product; following the tag there is a motif for thrombin protease cleavage. In the pET expression system, transcription of the recombinant gene is performed by the bacteriophage T7 RNA polymerase[78]. The expression hosts are lysogens of bacteriophage λDE3. These lysogens have a chromosomal copy of the T7 RNA polymerase gene under the control of the lacUV5 promoter. In the absence of the disaccharide allolactose (β-Gal-(1,6)-Glc) or its non-hydrolysable analog isopropyl-β-D-thiogalactopyranoside (IPTG), the lacI repressor protein is bound to the lacUV5 promoter inhibiting transcription. Binding of allolactose or IPTG reduces the affinity of the lacI repressor for DNA and allows transcription to proceed. The transcription in the expression plasmid is also control via a lac operator. To further suppress recombinant protein expression in the absence of IPTG, the pET15b plasmid also carries the coding sequence for the lacI repressor protein.

*Construction of the E. coli expression plasmid for hCD5d3*

The DNA encoding the whole extracellular domain of *h*CD5 was a kind gift from Dr Marion H. Brown and Prof Neil A. Barclay at the MRC Cellular Immunology Unit, Sir William Dunn School of Pathology, Oxford.

The system for prokaryotic expression chosen was ProEX HT (Invitrogen). The vector incorporates an N-terminal hexa-histidine sequence, followed by a tobacco etch virus (TEV) protease cleavage site. Transcription is performed by the endogenous bacterial

[2] Full details of the protocols are provided in Appendix A.

RNA polymerase and initiated via the strong Trc promoter. The vector also carries a lacIq repressor gene to suppress expression in the absence of IPTG.

The section of *h*CD5 comprising residues 269 to 374 was amplified by PCR (one cycle: 94°C for 5 min; 30 cycles: 95°C for 30 s, 55°C for 45 s, 72°C for 2 min; one cycle: 72°C for 5 min) using Pfu polymerase (Stratagene), to introduce an NcoI restriction enzyme site at the 5' end and a stop codon followed by an XhoI recognition site at the 3' end. The primers used were the following:

Forward 5'-GAT GAT GAT GAT ccatgg GT TTC CAG CCC AAG GTG-3'
Reverse 5'-GAT GAT GAT GAT ctcgag CTA GCC TGC GGG GTT TGG ATC-3'

The PCR product was purified using the QIAquick PCR purification kit (Qiagen). The vector and the purified PCR product were digested at the same time with NcoI and XhoI and purified by gel extraction using the QIAquick Gel extraction kit (Qiagen). Ligation was performed overnight at 14°C using T4 DNA ligase (NEB). The ligation reactions were used to transform subcloning efficiency DH5α competent cells (Invitrogen). Colonies were screened for the construct by colony PCR and restriction digestion. The sequence of the construct (Figure II.5b) was confirmed by automated DNA sequencing using dRhodamine dye-terminator chemistry (Applied Biosciences). The sequencing reaction products were run on a Abi Prism 377 at the Ludwig Institute of Cancer Research by Dr Athena Nikitopoulou. The calculated properties of the protein product based on their amino acid sequence, as computed by Expasy's ProtParam appear in Table II.8.


**Figure II.5** Amino acid sequence of the constructs for the N-terminal (*h*CD5d1) and C-terminal (*h*CD5d3) SRCR domains of human leukocyte receptor CD5. The shaded regions correspond to the SRCR domain sequence.

**Table II.8 Predicted properties of constructs**

| | *h* CD5d1 | *h* CD5d3 |
|---|---|---|
| Number of residues | 131 | 132 |
| Molecular weight [kDa] | 14.7 kDa | 14.9 kDa |
| Extinction coefficient* at 280 nm (reduced) [$M^{-1}cm^{-1}$] | 27880 | 24750 |
| Extinction coefficient* at 280 nm (oxidised) [$M^{-1}cm^{-1}$] | 28360 | 25230 |
| Theoretical pI | 8.20 | 6.42 |

* Calculated with the Gill and von Hippel equation.
Gill,SC & von Hippel,PH (1989)*Anal. Biochem.* 182(2):319-316

*Over expression in Luria-Bertani medium*

*E. coli* BL21(DE3) cells were transformed with the expression construct, plated onto Petri dishes with LB/agar and 50 μg/ml carbenicillin, and grown overnight at 37°C. A single colony was used to inoculate 10 ml of LB medium with 50 μg/ml carbenicillin, and the culture was incubated overnight at room temperature with shaking. Each of two 500 ml aliquots of LB medium with 50 μg/ml carbenicillin in a 2-litre baffled flask was inoculated with 1 ml of the overnight culture. The cultures were grown at 37°C with vigorous shaking until the $OD_{600}$ reached 1.0. Protein expression was then induced by addition of IPTG to a final concentration of 1 mM. Cells were harvested 4-12 hours after induction by centrifugation at 6000 g for 20 min.

*Over expression in $^{15}N$-labelled minimal media for $^{1}H/^{15}N$ 2D NMR*

A single freshly transformed BL21(DE3) colony was used to inoculate 10 ml of LB medium with 50μg/ml carbenicillin. The culture was incubated at 37°C with shaking for eight hours. One ml of this culture was used to inoculate 100 ml of minimal medium (6 g/l $Na_2HPO_4$, 3 g/l $KH_2PO_4$, 0.5 g/l NaCl, 1 g/l $(^{15}NH_4)_2SO_4$, 2 g/l D-glucose, metal salts and vitamins) with 50 μg/ml carbenicillin in a 500 ml baffled flask. The 100 ml culture was incubated overnight at 30°C with vigorous shaking. Cells were harvested by centrifugation at 4000 g for 30 min and resuspended in 50 ml of fresh minimal medium. Each one of four 2-litre baffled flasks with 500 ml of $^{15}N$-labelled minimal medium with 50 μg/ml carbenicillin was inoculated with 5 ml of the cell suspension. The flasks were incubated at 37°C with vigorous shaking for 8-12 hrs. Cells were harvested by centrifugation at 6000 g for 20 min.

*Cell disruption, and isolation and solubilisation of inclusion bodies*

Cell pellets were resuspended in a 50 mM potassium phosphate buffer pH 7.0 with 0.1 M NaCl and disrupted using a French pressure cell hydraulic press. The lysate was then centrifuged for 30 minutes at 27,000 g. The resulting pellet was resuspended using a mortar and pestle type homogeniser in a 50 mM potassium phosphate buffer pH 7.0 with 0.1 M NaCl and 2 M urea and incubated for 15 minutes at room temperature. The washed inclusion bodies were isolated by centrifugation at 27,000 g for 30 minutes. The pellet was resuspended, using a mortar and pestle type homogeniser, in 0.1 M potassium phosphate, 10 mM Tris-HCl, 6 M GdnHCl pH 8.0, and then incubated at room

temperature for 2 hrs with gentle stirring. The suspension was clarified by centrifugation at 27,000 g for 60 minutes.

*Purification of the expression product by immobilised metal ion affinity chromatography*

Gravity-driven chromatography was carried out in denaturing conditions using Ni-imminodiacetic acid (IDA) agarose (His-Bind, Novagen). The resin was equilibrated at pH 8.0 with 5 column volumes (CV) of the same buffer used to solubilise the inclusion bodies. The inclusion bodies obtained from a 0.5 litre of culture were purified as a single load onto 5 ml of resin. After loading the sample the column was washed with 10 CV of a buffer with the same composition but pH 6.0. The recombinant protein was eluted with 4 CV of the buffer at pH 4.5.

*Reduction of cysteines and isolation of the monomeric species by size exclusion chromatography*

The eluate was adjusted to pH 8.5 by drop-wise addition of 1 M NaOH and concentrated to 4 ml by centrifugal ultrafiltration (Vivaspin 20, Vivascience). Solid DTT (Melford Labs) was added to achieve a final concentration of 0.1 M and the solution was incubated overnight at RT with gentle stirring. SEC was performed on a 55 x 1.6 cm Superdex 75 column (Amersham Biosciences). The column was equilibrated with a 0.1 M potassium phosphate buffer pH 4.5, 10 mM Tris-HCl, 6 M GdnHCl (Melford Labs). Fractions corresponding to the protein monomer were pooled and concentrated by centrifugal ultrafiltration to yield solutions above 10 mg/ml as measured by absorbance at 280 nm.

*Preparation of denatured RNAase A and lysozyme*

Lyophilised hen egg white lysozyme and bovine pancreatic ribonuclease A (RNase A) (both Sigma) were resuspended in a 0.1 M potassium phosphate, 10 mM Tris-HCl buffer pH 8.0 with 6 M GdnHCl to a final concentration around 10 mg/ml. Solid DTT was added to obtain a final concentration of 0.1 M. Solutions were incubated overnight at 4°C with gentle stirring. Protein concentration was estimated by absorbance at 280 nm (Reduced lysozyme: $\varepsilon$=37980 $M^{-1}cm^{-1}$ and 14.3 kDa M.W; Reduced RNase A $\varepsilon$=7680 $M^{-1}cm^{-1}$ and M.W. 13.6 kDa).

*Design of the oxidative folding screen*

Based on bibliographic research and previous experience, the parameters selected for the screen were: temperature, pH, protein concentration and the presence of three additives, arginine, GdnHCl and glycerol. Each of these factors was assessed at two arbitrary values (levels). The values chosen for temperature were 4°C and RT and for pH were 8.0 and 9.0. Arginine was screened at 0 and 0.6 M, GdnHCl at 0 and 0.5 M, and glycerol at 0 and 20% (m/v). Final protein concentration was assessed at 0.1 mg/ml and 0.5 mg/ml for lysozyme and RNase A, and 0.1 mg/ml and 0.2 mg/ml for the *h*CD5d1 and *h*CD5d3 constructs.

The experiment design is a two-level one-eighth fractional factorial that assesses six parameters using eight combinations of levels for such parameters. The design was constructed following standard statistical procedures[77,79]. Briefly, the levels for the first three variables appear in every possible combination at two levels, *i.e.* $2^3= 8$ combinations. The levels for the fourth factor are the product of the levels for the first and the second factors (*e.g.* In B1 the level for the first factor is - and the level for the second factor is -, the result of the multiplication is +, and this is the level for the fourth factor). In the same manner, the levels for the fifth factor are generated by multiplying the levels for the first and third factors, and the levels for the sixth factor are obtained by multiplying the levels for the second and third factors (Table II.9).

*Oxidative folding screen*

Denatured and reduced protein solutions were diluted to the appropriate concentration into 4-6 ml of degassed 0.1 M Tris HCl buffer with 0.5 mM cystine and 5 mM cysteine and each of the eight conditions listed in Table II.9. Samples were incubated with gentle stirring for 48 hrs. After centrifugation for 10 minutes at 4000 g, the 8 supernatants (3-5 ml of each) were dialysed together against 3.5 l of a 0.05 M potassium phosphate buffer pH 6.0 with 0.2 M NaCl and 1 mM EDTA.

**Table II.9 Experimental conditions for the *in vitro* oxidative protein folding screen**

| Buffer | Factors* | | | | | | Conditions | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | Arginine [M] | Gdn HCl [M] | Glycerol % | Temp | pH | Protein [mg/ml] |
| B1 | - | - | - | + | + | + | 0 | 0 | 0 | Room | 9.0 | 0.5 |
| B2 | + | - | - | - | - | + | 0.6 | 0 | 0 | 4 °C | 8.0 | 0.5 |
| B3 | - | + | - | - | + | - | 0 | 0.5 | 0 | 4 °C | 9.0 | 0.1 |
| B4 | + | + | - | + | - | - | 0.6 | 0.5 | 0 | Room | 8.0 | 0.1 |
| B5 | - | - | + | + | - | - | 0 | 0 | 20 | Room | 8.0 | 0.1 |
| B6 | + | - | + | - | + | - | 0.6 | 0 | 20 | 4 °C | 9.0 | 0.1 |
| B7 | - | + | + | - | - | + | 0 | 0.5 | 20 | 4 °C | 8.0 | 0.5 |
| B8 | + | + | + | + | + | + | 0.6 | 0.5 | 20 | Room | 9.0 | 0.5 |

* **A** Arginine concentration [M];**B** Guanidine concentration [M];**C** Glycerol concentration (w/v) [%];**D** Temperature; **E** pH; **F** Protein concentration [mg/ml]

***Reversed-phase HPLC***

Each protein solution (1 ml for lysozyme and RNase A; 2 ml for *h*CD5d3 and 5 ml for *h*CD5d1) was loaded onto a Sephasil Protein C4 (300 Å, 5μm, 4.6 mm x 100 mm) column connected to an Äkta Purifier chromatography system (Amersham Biosciences). Solvent A was 5% (v/v) acetonitrile with 0.1% (v/v) trifluoroacetic acid (TFA) and solvent B was 95% acetonitrile 0.1% TFA. After injection, unbound material was eluted with 100% Solvent A with a flow rate of 2 ml/min for 2 CV. Protein was then eluted with a linear gradient from 0 to 45% Solvent B within 9 CV. Protein concentration was monitored by measuring absorbance at 220 and 280 nm.

*Peak integration*

Peak area was calculated using the Unicorn 4.0 software package (Amersham Biosciences). The elution volume range corresponding to each peak was set manually and was the same for all runs. Baseline calculation was performed automatically using the morphological algorithm implemented in the Unicorn package.

*Non-denaturing polyacrylamide gel electrophoresis for basic proteins*

The protocol used is an adaptation of the method for non-denaturing polyacrylamide gel electrophoresis (PAGE) for basic proteins described by D.P. Goldenberg[80]. The main difference was that ammonium persulphate (APS) was used instead of riboflavin to catalyse the polymerisation of acrylamide. APS can potentially oxidise the loaded protein samples and it is important to remove it once the gel has cast. In order to remove the APS gels were then pre-run at 30 mA for 2 hrs. The separating gel was 15% acrylamide, 0.4% bisacrylamide with an acetate buffer pH 4.0, and the stacking gel was 2.5% acrylamide, 0.3% bisacrylamide with an acetate buffer pH 5.0. Gels were 1.5 mm thick in order to prevent band distortion. The electrode buffer was 0.4 M β-alanine pH 4.0. After pre-running, tracking dye (50% glycerol, 0.2% methyl green) was added to the samples before loading. Electrophoresis was carried out at 10 mA for 3 hrs. Gels were stained with Coomassie blue.

*Preparative folding and purification of hCD5d1*

The inclusion bodies obtained from a 2-litre culture were purified and reduced, as described for the folding screen, and then diluted into 200 ml of the buffer for run B6 (0.1 M Tris pH 9.0 with 0.6 M arginine, 20% glycerol, 5 mM cysteine, 0.5 mM cystine at 4°C). Unfolded and reduced protein was added to the folding buffer in up to 4 steps, one addition every 2 hrs[81]. Each of the additions increased the protein concentration by 0.05 mg/ml. 48 hrs after the last addition, the protein solution was concentrated to an approximate final volume of 30 ml using a ultrafiltration stirred cell with a membrane with a molecular weight cut-off of 3kDa (Amicon, Millipore). Extensive dialysis to remove the arginine was performed against 0.05 M potassium phosphate buffer pH 6.0 with 0.2 M NaCl, 1 mM EDTA and 0.1% $NaN_3$. The solution was concentrated to ~20 ml using the stirred cell and then to ~1.5 ml by centrifugal ultrafiltration (Vivaspin 20, Vivascience), then loaded onto the Superdex 75 SEC column equilibrated in the same potassium phosphate buffer. Fractions containing the monomer were pooled and concentrated to ~0.5 ml.

*Preparative folding and purification of hCD5d3*

The methodology was very similar to that used for *h*CD5d1 except that the folding buffer was that used for run B4 (0.1 M Tris pH 8.0 with 0.6 M arginine, 0.5 M Gdn HCl, 5 mM cysteine, 0.5 mM cystine at RT).

*Nuclear magnetic resonance spectroscopy*

Native and *in vitro* folded proteins were dissolved in 0.05M potassium phosphate buffer pH 6.0 with 0.2 M NaCl, 1 mM EDTA and 0.1% $NaN_3$. Unfolded and reduced samples were dissolved in 0.05M potassium phosphate buffer pH 8.0 with 6 M Gdn HCl and 0.1 M DTT.

NMR spectra were acquired by Dr Richard Harris on either a Varian UnityPlus or UnityINOVA spectrometer, operating at nominal $^{1}H$ frequencies of 500 MHz and 600 MHz respectively. 1D $^{1}H$ spectra were recorded with 4096 complex data points with a spectral width of 10 kHz. The 2D $^{1}H$-$^{15}N$ HSQC of *h*CD5d1 was recorded with 256 complex points and a spectral width of 4 kHz in the $^{1}H$ dimension and 242 complex points and a spectral width of 1,600 Hz in the $^{15}N$ dimension. For *h*CD5d3, the 2D $^{1}H$-$^{15}N$ HSQC was recorded with 1024 complex points and a spectral width of 4 kHz in the $^{1}H$ dimension and in the $^{15}N$ dimension 192 complex points were recorded with a spectral width of 1.6 kHz. Chemical shifts were indirectly referenced to sodium 2,2-dimethyl-2-silane-pentane-5-sulphonate (DSS) using absolute frequency ratios for the $^{1}H$ signals [82]. NMR data were processed using NMRpipe/NMRDraw[83]. The figures of the 2D experiments (Figure II.12) where made with the Plot2 program of AZARA[84].

## Results

*Analysis of the data obtained from the folding screen*

The unfolded and reduced proteins were set to fold by diluting them into the eight different conditions of the screen (Table 9). After 48 hours, the samples were dialysed against 50 mM potassium phosphate buffer pH 6.0, 0.2 M NaCl, 1 mM EDTA, 0.1% $NaN_3$ (w/v). Some of the folding buffers underwent considerable osmosis when dialysed because of the high concentration of arginine and/or GdnHCl, as a result, the volume of most of the runs increased. Dialysed samples were injected into a C4 RP-HPLC column and eluted with an increasing linear gradient of acetonitrile. Peak areas were computed and used as the measure for protein quantity. In order to compare folding yields among runs, peak areas were normalised taking into account the changes in volume (V) as well as the difference in nominal initial protein concentration, these normalisation factors appear in the tables of results. The same volume of protein solution for each of the four proteins was always injected into the column, however if the concentration of protein was so high as to saturate the detector, dilutions of the dialysed samples were injected in order to obtain accurate data. To account for deliberate dilution and changes in volume during dialysis, the formula to obtain the normalisation factor was:

$$\text{Normalisation factor} = \frac{V_{\text{after dialysis}} \times V_{\text{after dilution}}}{V_{\text{before dialysis}} \times C_{\text{initial}} \times V_{\text{before dilution}}}$$

The calculation of effects was performed as explained in the Introduction. Every factor, appears in its + level in four of the eight conditions, and in the other four, it appears in its – level. The value for the effect is then one quarter of the sum of the results for the runs in which the factor is its + level, minus one quarter of the sum of the results for the runs in which the factor is in its – level:

$$\text{Effect of X} = \frac{\sum \text{runs with X(+)} - \sum \text{runs with X(-)}}{4}$$

For the model proteins, a rough estimation of the percentage yield was also obtained. In order to estimate the amount of refolded protein present, a known mass of native protein was injected onto the RP-HLPC column and the area of the elution peak was compared to the area of the peaks obtained in the *in vitro* folding experiments.

*Oxidative folding screen for RNase A*

Bovine pancreatic ribonuclease A (EC 3.1.27.5) is probably the best studied protein in *in vitro* folding experiments. It is a monomeric 13.7 kDa secreted enzyme with four disulphide bonds and a pI of 9.45[85]. RNase A was able to fold in all of the screen conditions with yields of between 40% and 80%. The best conditions were in runs B3, B4, B6 and B8 with yields between 70% and 80%. (Figure II.6; Table II.10). The only parameter that appeared to reduce the yield was high protein concentration. On the other hand, the presence of both arginine and guanidine hydrochloride had a positive influence on the yield. There is also a slight positive effect for high pH. The influences of glycerol and temperature are not distinguishable from the experimental error.

*Oxidative folding screen for lysozyme*

Hen egg white lysozyme (EC 3.2.1.17 is a secreted 14.4 kDa monomeric protein with four disulphide bonds and a pI of 11.0[86]. Lysozyme refolded in buffers B3, B4, B6 and B8, with yields between 20% and 67%. The analysis of the effects showed that the detrimental factors to folding were high protein concentration and higher temperature while the presence of arginine and a pH value of 9.0 had a positive effect on the yield, as did GdnHCl but to a lesser extent. The influence of glycerol was negligible (Figure II.7; Table II.11).

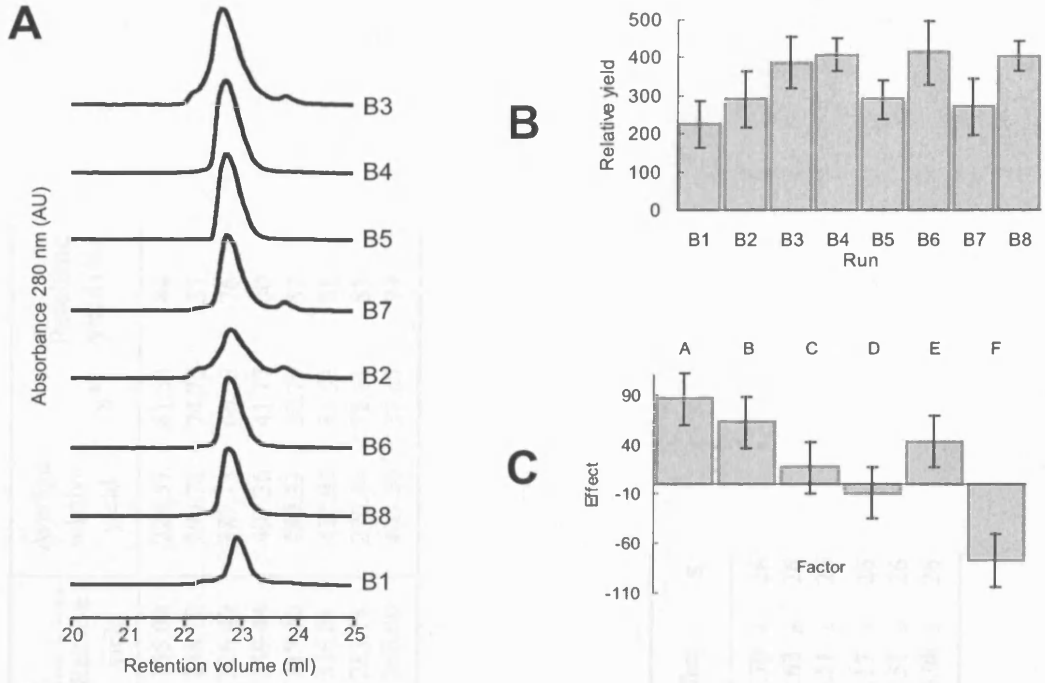


**Figure II.6** Results of oxidative refolding screen for RNase A.
(A) RP-HPLC analysis for quantifing refolding yield for different refolding conditions. Ordered by increasing area under the curve.
(B) Normalised area under the curve of the RP-HPLC chromatogram for each of the eight sets of conditions (Table II.10a).
(C) Main effects of each of the six factors (data in Table II.10b).

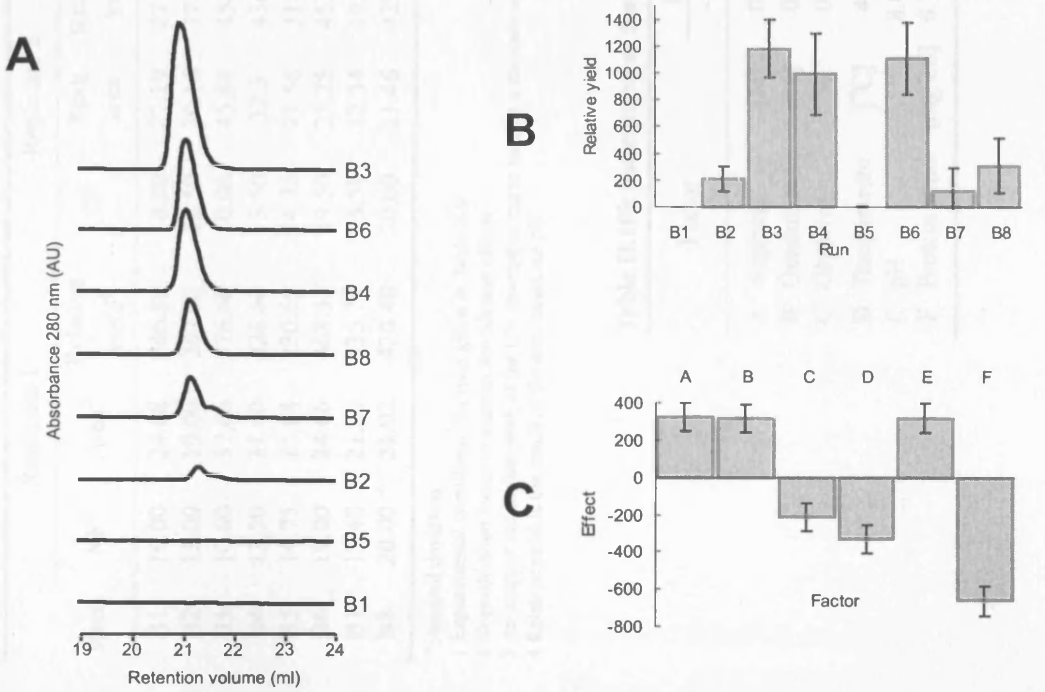


**Figure II.7** Results of oxidative refolding screen for lysozyme.
(A) RP-HPLC analysis for quantifing refolding yield for different refolding conditions. Ordered by increasing area under the curve.
(B) Normalised area under the curve of the RP-HPLC chromatogram for each of the eight sets of conditions (Table II.11a).
(C) Main effects of each of the six factors (data in Table II.11b).

**Table II.10a Oxidative refolding screen results for RNase A**

| Run[1] | Replicate 1 | | | Replicate 2 | | | Replicate 3 | | | Average relative yield | s* | Refolding yield (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NF[2] | Area[3] | Relative yield[4] | NF | Peak area | Relative yield | NF | Peak area | Relative yield | | | |
| B1 | 10.00 | 24.68 | 246.80 | 10.00 | 27.19 | 271.90 | 10.00 | 15.50 | 155.00 | 224.57 | 61.54 | 44 |
| B2 | 13.00 | 19.00 | 247.00 | 12.50 | 30.16 | 377.00 | 13.00 | 19.09 | 248.17 | 290.72 | 74.72 | 57 |
| B3 | 10.00 | 37.66 | 376.60 | 10.00 | 45.84 | 458.40 | 10.00 | 32.64 | 326.40 | 387.13 | 66.63 | 76 |
| B4 | 13.50 | 31.60 | 426.60 | 13.50 | 32.3 | 436.05 | 13.20 | 27.23 | 359.44 | 407.36 | 41.77 | 80 |
| B5 | 14.75 | 15.64 | 230.69 | 14.75 | 21.56 | 318.01 | 14.60 | 21.87 | 319.30 | 289.33 | 50.79 | 57 |
| B6 | 19.00 | 24.66 | 468.54 | 19.50 | 23.25 | 453.38 | 17.00 | 18.64 | 316.88 | 412.93 | 83.53 | 81 |
| B7 | 15.40 | 21.80 | 335.72 | 15.58 | 12.34 | 192.20 | 13.80 | 20.52 | 283.18 | 270.36 | 72.61 | 53 |
| B8 | 20.00 | 21.02 | 420.40 | 20.00 | 21.46 | 429.20 | 20.00 | 18.00 | 360.00 | 403.20 | 37.67 | 79 |

*Standard deviation

1 Experimental conditions for runs given in Table II.9

2 Normalization factor to account for dilution effects

3 Area under the elution peak of the UV absorption curve used as a measure of protein quantity

4 Relative yield is the result of the area times the NF

**Table II.10b Estimation of main effects for RNase A**

| | Factor | | Levels - | Levels + | Alias | Effect | | s |
|---|---|---|---|---|---|---|---|---|
| A | Arginine | [M] | 0 | 0.6 | A | 85.70 | ± | 26 |
| B | Guadinine | [M] | 0 | 0.5 | B | 62.63 | ± | 26 |
| C | Glycerol | % | 0 | 20 | C | 16.51 | ± | 26 |
| D | Temperature | [°C] | 4 | RT | D + AB | -9.17 | ± | 26 |
| E | pH | | 8.0 | 9.0 | E + AC | 42.51 | ± | 26 |
| F | Protein conc. | [mg/ml] | 0.1 | 0.5 | F + BC | -76.98 | ± | 26 |

**Table II.11a Oxidative refolding screen results for lysozyme**

| Run[1] | Replicate 1 | | | Replicate 2 | | | Replicate 3 | | | Average relative yield | s | Refolding yield (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NF[2] | Area[3] | Relative yield[4] | NF | Peak area | Relative yield | NF | Peak area | Relative yield | | | |
| B1 | 10.00 | 0.00 | 0.00 | 10.00 | 0 | 0.00 | 10.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0 |
| B2 | 13.50 | 20.40 | 275.40 | 35.00 | 7.1 | 248.50 | 14.00 | 7.40 | 103.60 | 209.17 | 92.41 | 12 |
| B3 | 10.00 | 130.65 | 1306.50 | 10.00 | 130.38 | 1303.80 | 10.00 | 92.99 | 929.90 | 1180.07 | 216.65 | 67 |
| B4 | 13.75 | 75.13 | 1033.04 | 13.75 | 91.74 | 1261.43 | 14.20 | 46.52 | 660.58 | 985.02 | 303.29 | 56 |
| B5 | 15.25 | 0.00 | 0.00 | 15.75 | 0.43 | 6.77 | 15.40 | 0.17 | 2.62 | 3.13 | 3.42 | 0 |
| B6 | 17.50 | 70.00 | 1225.00 | 21.00 | 61.16 | 1284.36 | 17.80 | 44.73 | 796.19 | 1101.85 | 266.37 | 63 |
| B7 | 16.00 | 1.20 | 19.20 | 14.75 | 0.99 | 14.60 | 13.60 | 23.08 | 313.89 | 115.90 | 171.48 | 7 |
| B8 | 20.50 | 10.51 | 215.46 | 22.50 | 7.09 | 159.53 | 19.60 | 27.41 | 537.24 | 304.07 | 203.85 | 17 |

*Standard deviation
1 Experimental conditions for runs given in Table II.9
2 Normalization factor to account for dilution effects
3 Area under the elution peak of the UV absorption curve used as a measure of protein quantity
4 Relative yield is the result of the area times the NF

**Table II.11b Estimation of main effects for lysozyme**

| | Factor | | Levels - | Levels + | Alias | Effect | | s |
|---|---|---|---|---|---|---|---|---|
| A | Arginine | [M] | 0 | 0.6 | A | 325.25 | ± | 78 |
| B | Guadinine | [M] | 0 | 0.5 | B | 317.73 | ± | 78 |
| C | Glycerol | % | 0 | 20 | C | -212.32 | ± | 78 |
| D | Temperature | [°C] | 4 | RT | D + AB | -328.69 | ± | 78 |
| E | pH | | 8.0 | 9.0 | E + AC | 318.20 | ± | 78 |
| F | Protein conc. | [mg/ml] | 0.1 | 0.5 | F + BC | -660.23 | ± | 78 |

*Expression and purification of hCD5d1 and hCD5d3*

Both *h*CD5 constructs expressed at high levels in BL21(DE3) cells and accumulated exclusively as inclusion bodies (Figures II.8a and II.9a). The expression levels reached their maximum at 3-4 hrs. The inclusion bodies were solubilised using 6 M GdnHCl, the recombinant protein was purified using immobilised metal ion affinity chromatography (IMAC). Samples were reduced and the monomeric species were isolated and by size-exclusion chromatography (SEC). Figures II.8 and II.9 show typical purification results for each of the *h*CD5d1 and *h*CD53 protein products.

Early attempts to fold these two domains were made without isolating the monomer and the folding yields were significantly lower than those obtained after the introduction of the SEC step. The efficiency of the protein reduction step varied from batch to batch. Considerable amounts of oligomers and/or high molecular weight contaminants were present in some batches. These disulphide-bonded oligomers and/or contaminants may have been responsible for the overall low *in vitro* folding yields. It is generally accepted that the folding yields are affected by the presence of common contaminants such as other proteins, lypopolysaccharides and nucleic acids[63,87-90]. However, there is evidence that folding yields are not always affected by the presence of contaminants[91].

*Oxidative folding screen for hCD5d3 and hCD5d1*

After evaluating and tuning the methodology with the model proteins RNase and lysozyme, the folding screen was carried out for *h*CD5d1 and *h*CD5d3. The final protein concentrations after dialysis were much lower than for the model proteins. As a result, in the case of *h*CD5d3 twice the volume was required in order to detect protein in the RP-HPLC eluate and in the case of *h*CD5d1 five times the volume was necessary for detection.

Folded *h*CD5d3 was detected in all the screen conditions (Table II.13; Figure II.11). Conditions with arginine gave the highest yields, and the yield with run B4 was the best of all. The presence of arginine was by far the most important effect. Arginine and GdnHCl had a positive effect on the folding yield. Higher protein concentration was the most detrimental factor, whist the effect of glycerol and temperature was negligible.

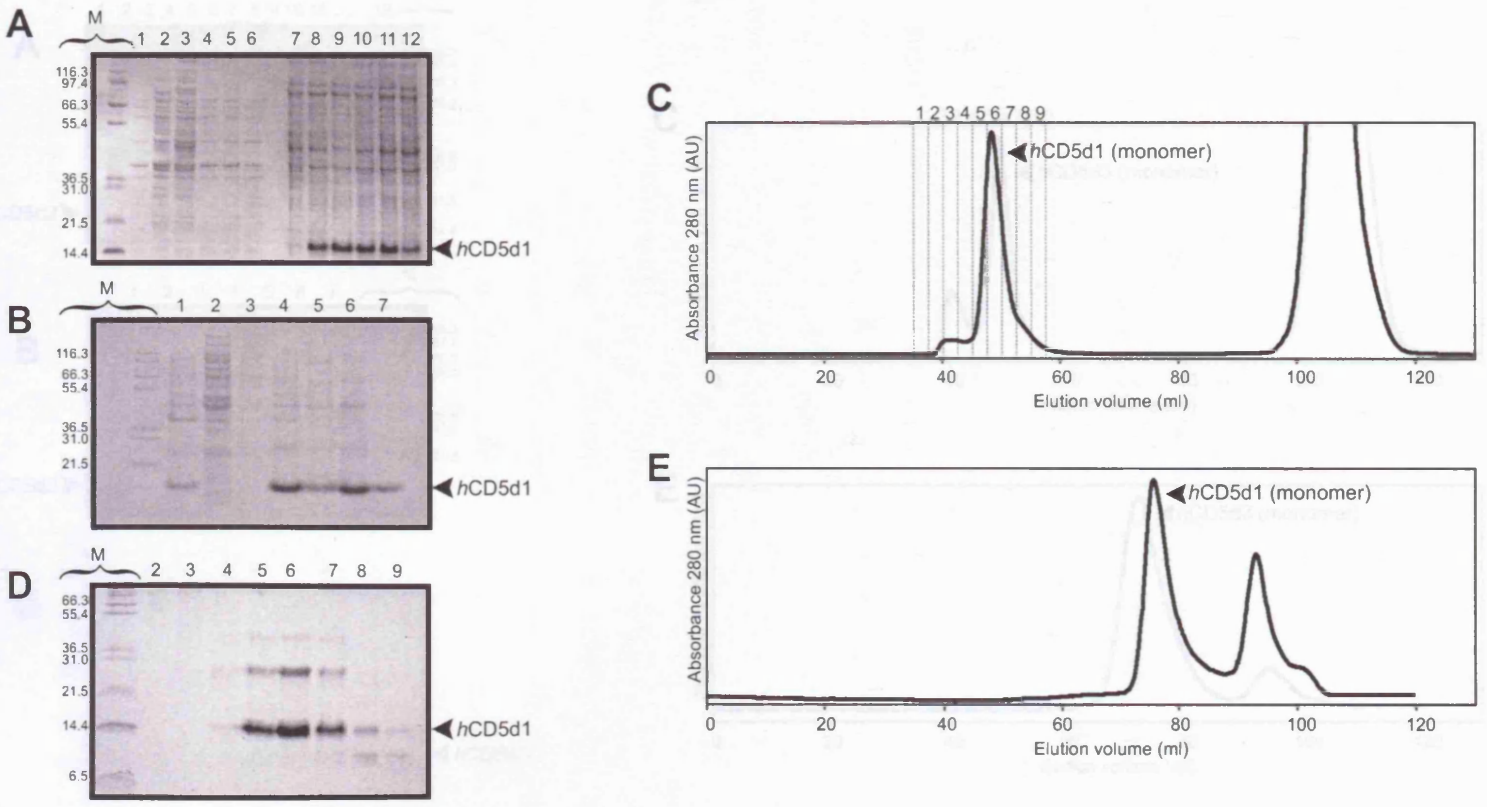


**Figure II.8** Expression and purification of *h*CD5d1.
(A) Reducing SDS-PAGE analysis of expression at different time points.
Lane M: molecular weight markers (kDa). Lanes 1-6: Fraction of the cell lysate soluble in 0.05 M phosphate, 0.2 M NaCl pH 7.0.
Lanes 7-12: Fraction soluble in 8 M urea, 0.1M phosphate, 0.01 M Tris pH 8.0.
Lanes:1 & 7, 0 min after induction; 2 & 8, 60 min; 3 & 9, 120 min; 4 &10, 180 min; 5 & 11, 240 min; 6 & 12, 960 min.
(B) Non-reducing SDS-PAGE analysis of purification of unfolded protein. Lanes: M, molecular weight markers (kDa); 1, total cell fraction; 2, soluble fraction; 3, 2 M urea wash; 4, 8 M urea soluble fraction; 5, eluate from Ni(II) IMAC; 6, reduced protein; 7, monomeric fraction from SEC.
(C) Size exclusion chromatography (SEC) in denaturing conditions in Superdex 75 matrix.
(D) Non-reducing SDS-PAGE analysis of fractions from denaturing SEC (panel C). Lane M: molecular weight markers (kDa). Other lane numbers correspond with fraction number.

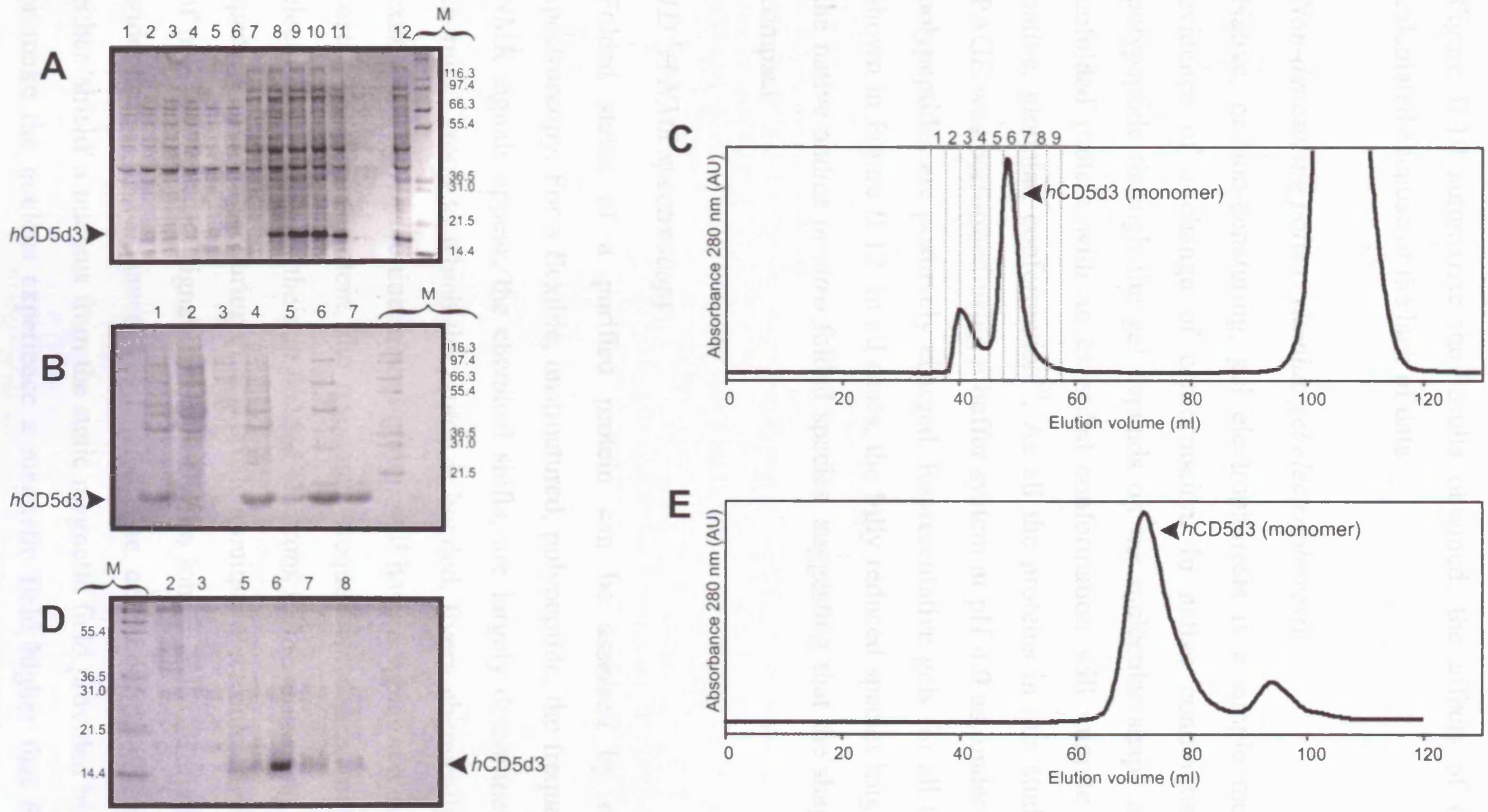


**Figure II.9** Expression and purification of *h*CD5d3.
(A) Reducing SDS-PAGE analysis of expression at different time points.
Lane M: molecular weight markers (kDa). Lanes 1-6: Fraction of the cell lysate soluble in 0.05 M phosphate, 0.2 M NaCl pH 7.0. Lanes 7-12: Fractions soluble in 8 M urea, 0.1 M phosphate, 0.01 M Tris pH 8.0. Lanes: 1 & 7, 0 min after induction; 2 & 8, 60 min; 3 & 9, 120 min; 4 & 10, 180 min; 5 & 11, 240 min; 6 & 12, 960 min.
(B) Non-reducing SDS-PAGE analysis of purification of unfolded protein.
Lanes: M, molecular weight markers (kDa); 1, total cell fraction; 2, soluble fraction; 3, 2 M urea wash; 4, 8 M urea soluble fraction; 5, eluate from Ni(II) IMAC; 6, reduced protein; 7, monomeric fraction from SEC.
(C) Size exclusion chromatography (SEC) in denaturing conditions in Superdex 75 matrix.
(D) Non-reducing SDS-PAGE analysis of fractions from denaturing SEC (panel C). Lane M: molecular weight markers (kDa). Other lane numbers correspond with fraction number.

In contrast to the model proteins and *h*CD5d3, *h*CD5d1 immediately precipitated upon dilution into some of the screen buffers. This straightforward qualitative observation was sufficient to indicate conditions that were not suitable for folding. *h*CD5d1 was only able to fold in two of the eight conditions and then with a very poor yield. Table II.12 and Figure II.10 summarise the results obtained, the effects of the variables were not calculated because of the lack of data.

*Non-denaturing polyacrylamide gel electrophoresis*

Native, or non-denaturing, gel electrophoresis is a simple method that can provide evidence of a change of conformation. In native conditions, the mobility of the polypeptide through the gel depends on the molecular shape and the net charge. An unfolded protein with an extended conformation will migrate more slowly than the native, globular, conformation[80]. As all the proteins in this study are basic, the native PAGE was performed using a buffer system at pH 4.0 as, under these conditions, basic polypeptides are positively charged. Representative gels for all the studied proteins are shown in Figure II.12. In all cases, the fully reduced species migrated more slowly than the native and/or *in vitro* folded species, suggesting that the shape of the latter is more compact.

*1D $^{1}H$ NMR spectroscopy*

Folded status of a purified protein can be assessed by one-dimensional NMR spectroscopy. For a flexible, unstructured, polypeptide, the frequencies at which the $^{1}H$-NMR signals appear, the chemical shifts, are largely determined by the nature of the chemical group to which the proton is bonded. Every chemically equivalent group, for example, every threonine methyl group, will have a signal at a similar frequency. In the case of a folded protein, the resonance frequencies depend not only on the bonding electrons but also on the non-bonded electrons of the atoms that are physically close in space. These neighbouring groups create a unique electronic environment for each group of protons, thus the signals are typically no longer equivalent and are dispersed over a wider range of resonance frequencies. The neighbouring electronic environment can either 'shield' a nucleus from the static magnetic field provided by the spectrometer ($B_0$), or make the nucleus experience a magnetic field higher that $B_0$ (deshielding). If the nucleus is shielded the resonance frequency is lower and the signal shifts to the right of the spectrum. If it is deshielded the resonance frequency is higher and the peak appears to

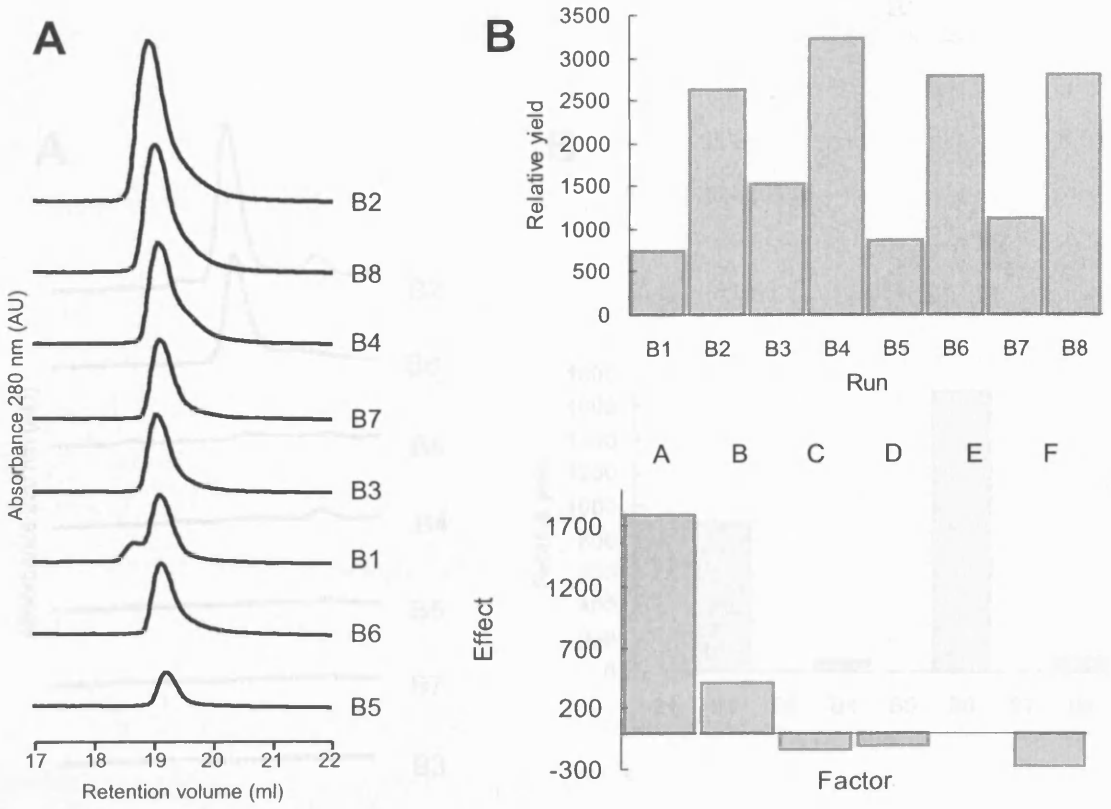


**Figure II.10** Results of oxidative refolding screen for *h*CD5d3.
(A) RP-HPLC analysis for quantifing folding yield for different folding conditions, plotted in order of increasing area under the curve.
(B) Normalised area under the curve of the RP-HPLC chromatogram for each of the eight sets of folding conditions (Table II.12a).
(C) Main effects of each of the six factors (Table II.12b)

**Table II.12a Oxidative folding screen results for *h* CD5d3**

| Run[1] | NF[2] | Area[3] | Relative yield[4] |
|---|---|---|---|
| B1 | 25 | 29 | 733 |
| B2 | 32 | 82 | 2634 |
| B3 | 50 | 31 | 1526 |
| B4 | 68 | 47 | 3223 |
| B5 | 75 | 12 | 873 |
| B6 | 94 | 30 | 2773 |
| B7 | 36 | 32 | 1134 |
| B8 | 45 | 62 | 2795 |

1 Experimental conditions for runs given in Table II.9
2 Normalization factor to account for dilution effects
3 Area under the elution peak of the UV absorption curve used as a measure of protein quantity
4 Relative yield is the result of the area times the NF

**Table II.12b Estimation of main effects for *h* CD5d3**

| | Factor | | Levels | | Alias | Effect |
|---|---|---|---|---|---|---|
| | | | - | + | | |
| A | Arginine | [M] | 0 | 0.6 | A | 0 |
| B | GndHCl | [M] | 0 | 0.5 | B | 0 |
| C | Glycerol | % | 0 | 20 | C | 0 |
| D | Temperature | | 4°C | Room | D + AB | -111 |
| E | pH | | 8.0 | 9.0 | E + AC | -10 |
| F | Protein conc. | [mg/ml] | 0.1 | 0.5 | F + BC | -275 |

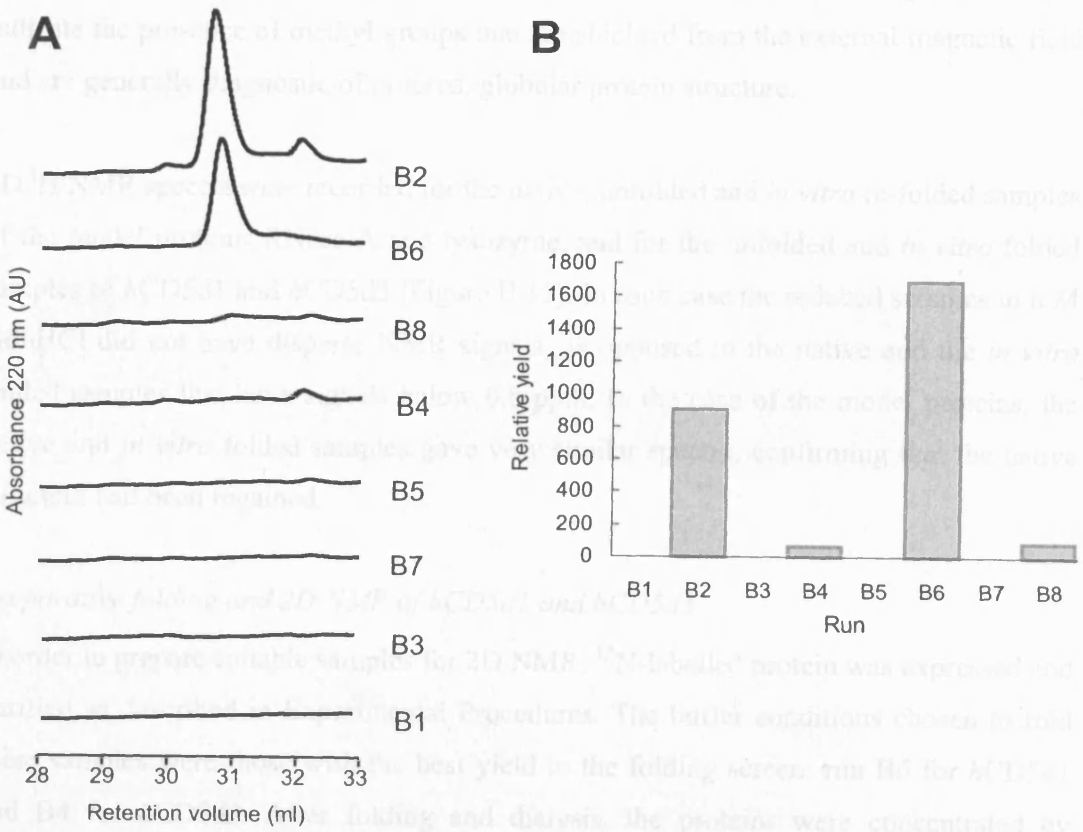


**Figure II.11** Results of oxidative refolding screen for *h*CD5d1.
(A) RP-HPLC analysis for quantifying folding yield for different folding conditions, plotted in order of increasing area under the curve.
(B) Normalised area under the curve of the RP-HPLC chromatogram for each of the eight sets of folding conditions (Table II.13).

**Table II.13 Oxidative folding screen results for *h* CD5d1**

| Run[1] | NF[2] | Area[3] | Relative yield[4] |
|---|---|---|---|
| B1 | 5.0 | 0 | 0 |
| B2 | 13.2 | 69 | 907 |
| B3 | 20.0 | 0 | 0 |
| B4 | 27.6 | 2 | 62 |
| B5 | 28.8 | 0 | 0 |
| B6 | 36.0 | 47 | 1694 |
| B7 | 14.8 | 0 | 0 |
| B8 | 18.0 | 5 | 90 |

1 Experimental conditions for runs given in Table II.9
2 Normalization factor to account for dilution effects
3 Area under the elution peak of the UV absorption curve used as a measure of protein quantity
4 Relative yield is the result of the area times the NF

the left of the spectrum. In a flexible unstructured polypeptide, amino acid methyl groups display resonance frequencies between 2.1 and 0.8 ppm[92]. Signals at the right of 0.8 ppm indicate the presence of methyl groups that are shielded from the external magnetic field and are generally diagnostic of ordered, globular protein structure.

1D $^{1}$H NMR spectra were recorded for the native, unfolded and *in vitro* re-folded samples of the model proteins RNase A and lysozyme, and for the unfolded and *in vitro* folded samples of *h*CD5d1 and *h*CD5d3 (Figure II.12). In each case the reduced samples in 6 M GdnHCl did not have disperse NMR signals, as opposed to the native and the *in vitro* folded samples that have signals below 0.8 ppm. In the case of the model proteins, the native and *in vitro* folded samples gave very similar spectra, confirming that the native structure had been regained.

*Preparative folding and 2D NMR of hCD5d1 and hCD5d3*

In order to prepare suitable samples for 2D NMR, $^{15}$N-labelled protein was expressed and purified as described in Experimental Procedures. The buffer conditions chosen to fold these samples were those with the best yield in the folding screen: run B6 for *h*CD5d1 and B4 for *h*CD5d3. After folding and dialysis, the proteins were concentrated by ultrafiltration and finally purified using SEC, in order to remove any oligomers and traces of the folding buffer additives. For both constructs, there was significant precipitation at final protein concentrations above 200 μM, indicating that the constructs were not very soluble in the final NMR buffer.

The $^{1}$H-$^{15}$N single quantum heteronuclear coherence (HSQC) spectrum of a protein is regarded, by NMR spectroscopists, as its 'fingerprint'. The $^{1}$H-$^{15}$N HSQC spectrum contains a signal for every proton bound to a nitrogen atom. In the ideal case, each amino acid residue, except proline and the N-terminus, yields one peak corresponding to its backbone amide group. Additionally, side chain N-H bonds amine of His, Trp, Arg, Gln and Asn also give rise to signals in the spectrum. In the $^{1}$H-$^{15}$N HSQC of an unfolded protein the backbone NH signals appear heavily overlapped, mainly between 8.0 and 8.5 ppm in the $^{1}$H dimension, and between 120 and 125 ppm in the $^{15}$N dimension. The $^{1}$H-$^{15}$N HSQC spectrum of a folded protein typically shows a wider spread of signal frequencies. For *h*CD5d1 and *h*CD5d3 the $^{1}$H-$^{15}$N HSQC spectrum displays a broad

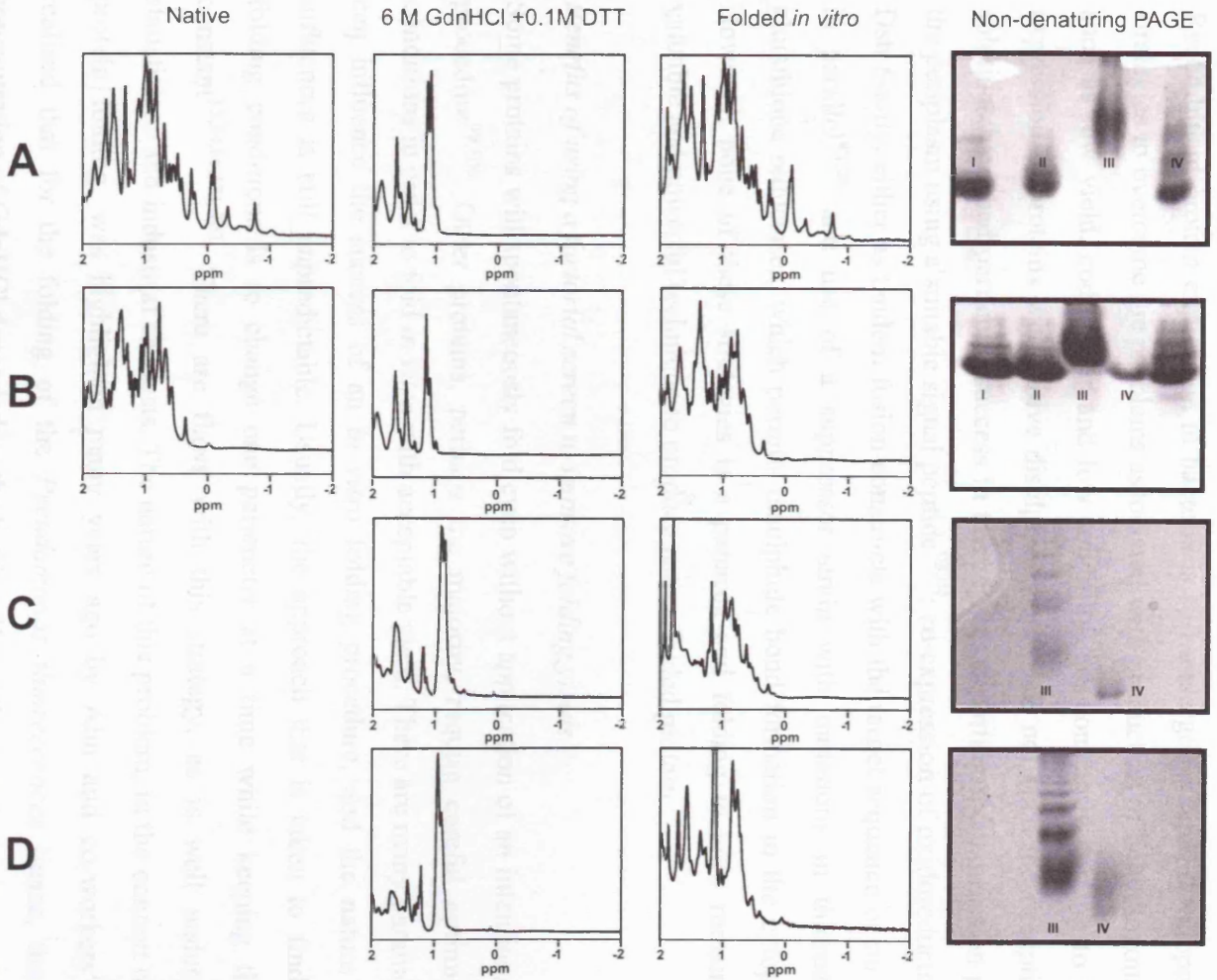


**Figure II.12** High field region of $^{1}$H 1D-NMR spectra and non-denaturing PAGE before and after *in vitro* folding.
(A) Lysozyme; (B) RNase A; (C) *h*CD5d1; (D) *h*CD5d3.
Native gel lanes: I. native; II. denatured with intact disulphide bonds; III. denatured with reduced disulphide bonds; IV. after *in vitro* folding.

range of NMR frequencies, indicative of the presence of a folded polypeptide chain (Figure II.13).

## Discussion

Recombinant protein expression in bacteria is still undergoing rapid development. New strategies to overcome the problems associated with production of heterologous proteins, such as low yield, codon bias and low solubility, are constantly coming to light. The expression of proteins with native disulphide bonds has not been the exception. Three solutions have had particular success in this regard: exporting the expression product to the periplasm using a suitable signal peptide[93,94]; co-expression of oxidoreductases of the Dsb family, either as tandem fusion constructs with the target sequence or on their own in parallel[95];[26] and use of a suppressor strain with mutations in thioredoxin and glutathione reductases, which permits disulphide bond formation in the cytoplasm[96-98]. However, none of these strategies is a panacea and folding *in vitro* remains a very valuable and powerful technique to produce native folded protein.

*Benefits of using a factorial screen to improve folding yields*

Some proteins will spontaneously fold even without application of an intentional folding procedure[99,100]. Other proteins, perhaps the majority, require careful optimisation of conditions in order to fold *in vitro* with acceptable yield. There are many parameters than can influence the success of an *in vitro* folding procedure, and the nature of these influences is still unpredictable. Usually, the approach that is taken to find suitable folding conditions is to change one parameter at a time while keeping the others constant[13,34,66,101-103]. There are flaws with this strategy, as is well understood by statisticians and industrial chemists. The nature of this problem, in the context of *in vitro* protein folding, was highlighted many years ago by Ahn and co-workers[104]. They realised that for the folding of the *Pseudomonas fluorescences* lipase, the optimal concentration of GdnHCl depended on the buffer pH and the protein concentration. Often the most favourable value for one factor varies when the value of another factor is changed. A one-factor-at-a-time strategy will probably fail if the optimal value for any of the variables involved depends on the value of another variable. The conclusions obtained from one-factor-at-a-time-experiments are not general, and they may not be

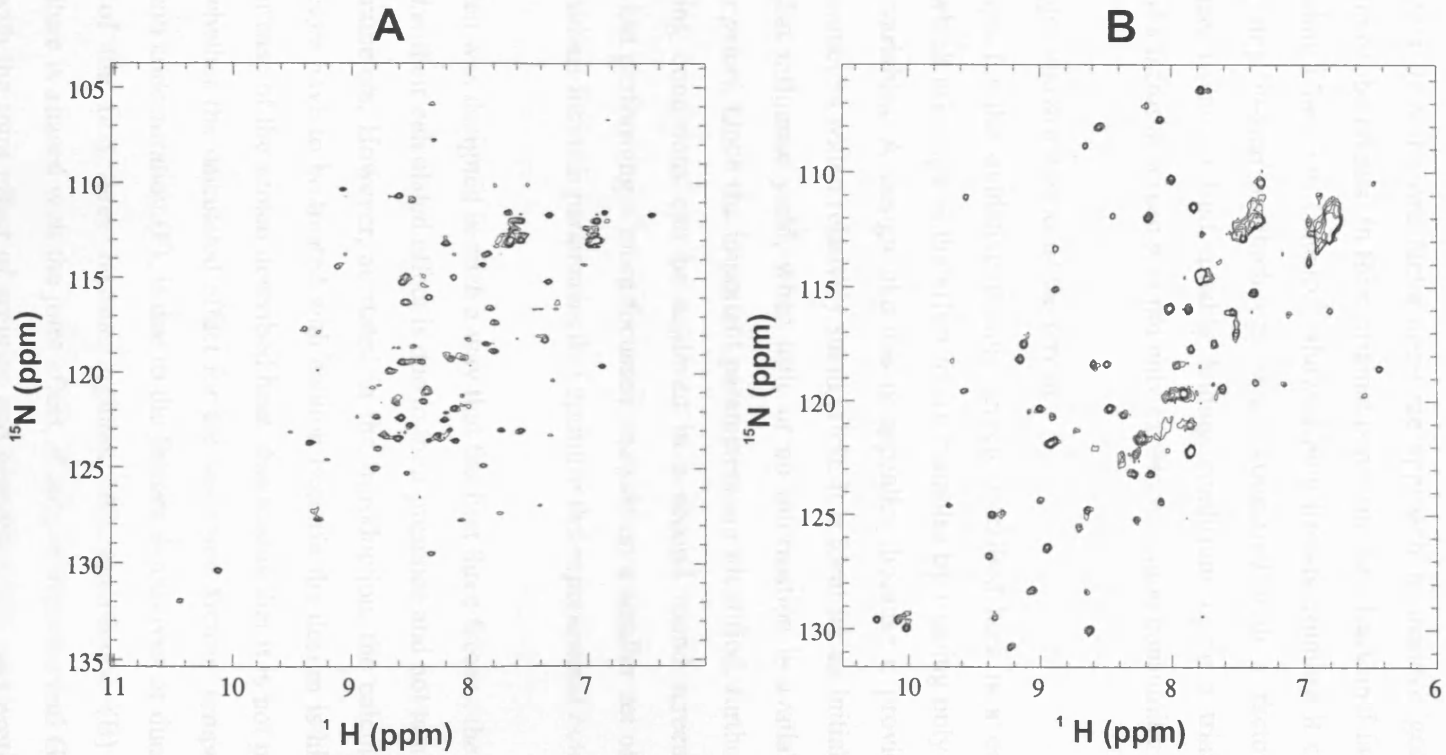


**Figure II.13**
(A) $^{1}$H-$^{15}$N HSQC spectrum of *h*CD5d1 at 278K in 50 mM potassium phosphate pH 6.0, 0.2 M NaCl, recorded at 500 MHz
(B) $^{1}$H-$^{15}$N HSQC spectrum of *h*CD5d3 at 278K in 50 mM potassium phosphate pH 8.0, 0.2 M NaCl, recorded at 600 MHz

reproduced if any of the other parameters is changed[77,79]. A factorial design is a better way to screen conditions efficiently and effectively. This strategy explicitly takes into account the possibility that the factors studied might not be independent of each other. Even if they are independent, the utility of a factorial design is that it requires fewer experiments than the one-factor-at-a-time approach to achieve good precision in the estimation of the effects. In their original paper on their fractional factorial screen for *in vitro* folding Chen and Gouaux[42] showed how time-consuming it can be to develop a 'rational step-by-step' methodology when compared with a factorial screen. In my experience, trying to find suitable folding conditions using a trial-an-error approach instead of a factorial screen was not only excessively time consuming but also fruitless.

*Advantages and drawbacks of the screen*

The design for the oxidative folding screen described here is a one-eighth fractional screen, which investigates the effect of six variables by running only eight combinations of such variables. A design like this is appealing because it provides information on many parameters with a relatively small effort. It is ideal for an initial screen for relevant factors that influence yield, when little or no information is available for a particular system *a priori*. Once the important parameters are identified, further improvements on the folding conditions can be achieved in a second round screen, also by factorial designs, but performing a more focussed analysis on a smaller set of factors in order to find the values for such parameters that optimise the experimental outcome.

The screen was designed in such a way that the first three factors, the additives, were not aliased, *i.e.* their calculated effect is due to their presence and not to a combination of the other parameters. However, as stated in the Introduction, the calculated effects for the other factors have to be treated with caution because the design is highly aliased. In the particular case of the screen described here, this means that it is not possible to know for certain whether the calculated effect for the last three factors: temperature (D), pH (E) and protein concentration (F), is due to the factors themselves or due to the combination of two of the first three factors: arginine (A), glutathione (B) and glycerol (C). Temperature is aliased with the joint effect, if any, of arginine and GdnHCl (AB); pH is aliased with the joint effect of arginine and glycerol (AC), and protein concentration is aliased with the joint effect of GdnHCl and glycerol (BC). If there evidence to support the existence of any of these joint effects, it should be investigated further. Heavy aliasing and the limited success in analysing individual effects does not negate the fact

that by this method the experimenter is able to find a set of conditions that produces natively folded protein with an acceptable yield, even if each and every parameter is not completely optimised.

A factorial screen should be regarded as a very malleable methodology. If the particular set of factors described here does not give the desired results, the values or the parameters to screen can be simply changed. Other parameters that might be tested are the ratio of redox shuffling reagents and the presence of PEG. The proposed levels and factors should not be regarded as permanently fixed, what is important is to become familiar with the theory and methodology of factorial designs, and use this knowledge to tailor the screen to suit particular needs. Once this technique has been set up and the experimenter has become dextrous in its implementation, it takes only one week from expressing a protein construct as inclusion bodies to having evidence that the construct is able to fold.

*Importance of general methods to characterise folding*

The efficient evaluation of different folding strategies and the ability of a great variety of additives to prevent aggregation requires a systematic method to quantify folded protein yields. This task is straightforward if the protein has a quantitative biological assay, such as catalytic activity, ligand binding or conformation-dependent antibody binding. When none of these methods exist for the target protein, or if many different proteins need to be tested, it is paramount to set up a robust general technique which allows not only to quantify, but also to distinguish between folded and unfolded protein. Chromatography and spectroscopy have proven to be useful as general detection methods for protein folding. The former methodology might be more attractive because it is readily available in any protein biochemistry laboratory. Previously two different chromatographic methods sensitive to changes in protein conformation have been reported to be applicable: analytical SEC and analytical HIC. The folding of disulphide-bonded proteins can also be assessed by analytical RPC. Both HIC and RPC are based on non-polar interactions between the matrix and the solvent-exposed groups in the protein. However, in RPC the interaction between the matrix and the protein is stronger than in HIC, and non-polar solvents are required for elution, while in HIC elution can be achieved using a decreasing gradient of salt concentration[105]. Proteins usually unfold in RPC conditions due to the presence of organic solvent and the acidic pH. However, if disulphide bonds hold the polypeptide chain in a compact state, the hydrophobic core is not completely

exposed upon denaturation without reduction, and the resulting conformation is still less hydrophobic than the fully unfolded species without the native disulphides. Evidence suggests that, after RPC disulphide-bonded proteins tend to refold readily once the organic solvent has been removed[4,106-108]. RP-HPLC has been used extensively to monitor protein folding in kinetic experiments and to identify the presence of the oxidised protein[26,109,110], and at least once to quantify folding yields[103]. Protein samples from folding experiments are often highly dilute, and protein cannot be detected in small loading volumes without prior concentration. In both HIC and RPC the sample volume is not a limiting factor as protein is retained and concentrated on the column matrix before elution conditions are applied. This feature is a big advantage for HIC and RPC over SEC, which requires small sample loading volumes. This aspect was crucial in the case of *h*CD5d1 and *h*CD5d3 whose folding yields using un-optimised conditions are very low. In order to detect any protein it was necessary to load 5-ml and 2-ml sample volumes, respectively, onto the analytical RP-HPLC column. In contrast, in applications of analytical SEC the loading volume should be limited to 200 μl.

When using chromatographic techniques for protein quantification it is important to ensure that the calculated peak areas truly reflect the amount of protein present in the samples. Peak shapes and baseline profiles greatly affect the accuracy of peak integration, and this issue was a major concern during this work, as the elution peaks from some runs had uneven profiles, most likely because there were intermediate folding species present. The problem was ameliorated by defining the elution time boundaries of the peaks manually and using the same boundaries for all the runs of the same protein, but this is a problem that requires more attention because it is vital for accurate quantification. An alternative to explore is to use maximum peak height as the quantifying parameter instead of peak area. The solvent gradient and the column matrix used could also be thoroughly optimised in order to obtain better resolution, however, this would probably make each measurement more time-consuming. It would also be valuable to ensure that there is a linear response between quantity of protein and peak area/height and to apply typical analytical chemistry strategies such as using an internal standard to account for unanticipated variations in the injection volume.

In the folding screen described in this project, non-denaturing gel electrophoresis was a useful method to detect changes in protein conformation. It is a simple and inexpensive technique and, as with chromatography, the necessary equipment is already found in

every biochemistry laboratory. The technique provides qualitative information on the degree of compactness of the polypeptide chain. An extended unfolded conformation will have reduced electrophoretic mobility compared to a folded globular entity. If the folded sample does migrate further than the unfolded protein, further investigation is justified, *e.g.* preparation of a sample for NMR experiments,

*Refolding of lysozyme and RNase A*

Model proteins provide a frame of reference when dealing with unfamiliar problems. However, the characteristic response that they show should be treated with reservation because they often represent the best case scenario. The wide range of conditions in which the two model proteins folded may correlate with the fact that they are secreted enzymes, evolutionary selected for their robustness, thermodynamic stability and solubility. RNase A was only marginally affected by the changes in the folding conditions and could refold to a greater or lesser extent in all buffers that were sampled (Figure II.6B). Lysozyme required low protein concentrations and arginine and/or GdnHCl to refold (Table II.11b). As expected, in both cases the most detrimental effect was high concentration of protein. The fact that high pH seemed to improve the folding yield has to be taken with caution, since this factor was aliased with the two-factor effect of arginine and glycerol, and arginine was the most beneficial factor (Tables II.10 and II.11)

*Folding of hCD5d1 and hCD5d3*

The oxidative folding screen was successful in rapidly finding suitable folding conditions for both *hCD5d1 and hCD5d3*, albeit with much lower yields than for the model proteins. The $^{1}H$-$^{15}N$ HSQC spectra proved that after *in vitro* folding, both constructs have, at least at some degree, a three dimensional structure as judged by the chemical shift dispersion of the NMR signals. Although the final concentration of protein obtained was enough to obtain a HSQC spectrum, at least five times higher concentration will be necessary to perform the NMR experiments needed for the assignment of the signals and structure determination. The low concentration could be a result of low folding yields or intrinsic low protein solubility. If the latter was the case, the composition of the NMR buffer needs to be optimised. It may be of benefit to study the use of fractional factorial screens to solve this particular problem.

To summarise, a novel screen to find suitable *in vitro* folding conditions for disulphide-bonded proteins was designed. The screen is based on a fractional factorial design that allows to systematically screen 6 different variables in 8 trial experiments. In each trial, *in vitro* folding is attempted by diluting the unfolded and reduced protein into a buffer consisting of a specific combination of such 6 variables. Folding yield is then assessed using RP-HPLC and non-denaturing gel electrophoresis. The screen was tested and optimised with RNase A and lysozyme and then used to find suitable *in vitro* folding conditions for two of the three extracellular domains of human leukocyte receptor CD5 (*h*CD5d1 and *h*CD5d3). NMR experiments confirmed that the *in vitro* folded *h*CD5d1 and *h*CD5d3 have a compact 3D structure. However, both CD5 domains appear to have intrinsic low solubility in the final NMR buffer and at this point are unsuitable for structure determination by NMR.

## References


1. Anson ML & Mirsky AE (1934) The equilibrium between active native trypsin and inactive denatured trypsin. *J Gen Physiol.* **17**, 393-398.
2. Mirsky AE & Anson ML (1929) Protein coagulation and its reversal: The reversal of the coagulation of haemoglobin. *J Gen Physiol.* **13**, 133-143.
3. Perutz MF (1980) Unboiling an egg (from 'Discovery', March 1940). In *Protein Folding. Proceedings of the 28th Conference of the German Biochemical Society.* (Jaenicke R ed.), Elsevier/North Holand Biomedical Press, Amsterdam.
4. Kroeff EP, Owens RA, Campbell EL, Johnson RD & Marks HI (1989) Production scale purification of biosynthetic human insulin by reversed-phase high-performance liquid chromatography. *J Chromatogr.* **461**, 45-61.
5. Mohanty AK, Mukhopadhyay UK, Grover S & Batish VK (1999) Bovine chymosin: production by rDNA technology and application in cheese manufacture. *Biotechnol Adv.* **17**, 205-217.
6. Rosenberg SA, Grimm EA, McGrogan M, Doyle M, Kawasaki E, Koths K & Mark DF (1984) Biological activity of recombinant human interleukin-2 produced in *Escherichia coli. Science.* **223**, 1412-1414.
7. Platis D & Foster GR (2003) High yield expression, refolding, and characterization of recombinant interferon α2/α8 hybrids in *Escherichia coli. Protein Expr Purif.* **31**, 222-230.

8. Rudolph R & Fischer S (1990) Process for obtaining renatured proteins. United States of America Patent number 4933434.

9. Orsini G & Goldberg ME (1978) The renaturation of reduced chymotrypsinogen A in guanidine HCl. Refolding versus aggregation. *J Biol Chem.* **253**, 3453-3458.

10. Maeda Y, Yamada H, Ueda T & Imoto T (1996) Effect of additives on the renaturation of reduced lysozyme in the presence of 4 M urea. *Protein Eng.* **9**, 461-465.

11. Hevehan DL & Clark ED (1997) Oxidative renaturation of lysozyme at high concentrations. *Biotechnol Bioeng.* **54**, 221-230.

12. Shi JH, Dong XY & Sun Y (2001) Effect of guanidinium chloride concentration on the renaturation of denatured lysozyme. *Acta Biochim Biophys Sinica.* **33**, 447-451.

13. Lu BY, Beck PJ & Chang JY (2001) Oxidative folding of murine prion mPrP(23-231). *Eur J Biochem.* **268**, 3767-3773.

14. Cleland JL, Builder SE, Swartz JR, Winkler M, Chang JY & Wang DIC (1992) Polyethylene glycol enhanced protein refolding. *Bio-Technology.* **10**, 1013-1019.

15. Cleland JL & Randolph TW (1992) Mechanism of polyethyleneglycol interaction with the molten globule folding intermediate of bovine carbonic anhydrase B. *J Biol Chem.* **267**, 3147-3153.

16. Ambrus A & Fesus L (2001) Polyethylene glycol enhanced refolding of the recombinant human tissue transglutaminase. *Prep Biochem Biotechnol.* **31**, 59-70.

17. Lilie H, Haehnel W, Rudolph R & Baumann U (2000) Folding of a synthetic parallel beta-roll protein. *FEBS Lett.* **470**, 173-177.

18. Nishiyama Y, Hardre-Lienard H, Miras S, Miege C, Block MA, Revah F, Joyard J & Marechal E (2003) Refolding from denatured inclusion bodies, purification to homogeneity and simplified assay of MGDG synthases from land plants. *Protein Expr Purif.* **31**, 79-87.

19. De la Cruz R, Whitby FG, Buczek O & Bulaj G (2003) Detergent-assisted oxidative folding of δ-conotoxins. *J Pept Res.* **61**, 202-212.

20. Chang SY, Tsai PC, Tseng CS & Liang PH (2001) Refolding and characterization of a yeast dehydrodolichyl diphosphate synthase overexpressed in *Escherichia coli. Protein Expr Purif.* **23**, 432-439.

21. Ayub MJ, Levin MJ & Aguilar CF (2001) Overexpression and refolding of the hydrophobic ribosomal P0 protein from *Trypanosoma cruzi*: a component of the P1/P2/P0 complex. *Protein Expr Purif.* **22**, 225-233.

22. Katzav-Gozansky T, Hanan E & Solomon B (1996) Effect of monoclonal antibodies in preventing carboxypeptidase A aggregation. *Biotechnol Appl Biochem.* **23**, 227-230.

23. Bezsudnova EY, Zherdev AV, Ermolenko DN, Yakovleva IV, Sviridov VV, Popov VO & Dzantiev BB (2003) Studies of peroxidase refolding in the presence of specific antibodies. *Appl Biochem Microbiol.* **39**, 446-453.

24. Altamirano MM, Garcia C, Possani LD & Fersht AR (1999) Oxidative refolding chromatography: folding of the scorpion toxin Cn5. *Nat Biotechnol.* **17**, 187-191.

25. Altamirano MM, Golbik R, Zahn R, Buckle AM & Fersht AR (1997) Refolding chromatography with immobilized mini-chaperones. *Proc Natl Acad Sci USA.* **94**, 3576-3578.

26. Maskos K, Huber-Wunderlich M & Glockshuber R (2003) DsbA and DsbC-catalyzed oxidative folding of proteins with complex disulfide bridge patterns *in vitro* and *in vivo*. *J Mol Biol.* **325**, 495-513.

27. Karuppiah N & Sharma A (1995) Cyclodextrins as protein folding aids. *Biochem Biophys Res Commun.* **211**, 60-66.

28. Rozema D & Gellman SH (1996) Artificial chaperone-assisted refolding of carbonic anhydrase B. *J Biol Chem.* **271**, 3478-3487.

29. Daugherty DL, Rozema D, Hanson PE & Gellman SH (1998) Artificial chaperone-assisted refolding of citrate synthase. *J Biol Chem.* **273**, 33961-33971.

30. Machida S, Ogawa S, Xiaohua S, Takaha T, Fujii K & Hayashi K (2000) Cycloamylose as an efficient artificial chaperone for protein refolding. *FEBS Lett.* **486**, 131-135.

31. Odorzynski TW & Light A (1979) Refolding of the mixed disulfide of bovine trypsinogen and glutathione. *J Biol Chem.* **254**, 4291-4295.

32. Werner MH, Clore GM, Gronenborn AM, Kondoh A & Fisher RJ (1994) Refolding proteins by gel filtration chromatography. *FEBS Lett.* **345**, 125-130.

33. Saini DK, Pant N, Das TK & Tyagi JS (2002) Cloning, overexpression, purification, and matrix-assisted refolding of DevS (Rv 3132c) histidine protein kinase of *Mycobacterium tuberculosis*. *Protein Expr Purif.* **25**, 203-208.

34. Panagabko C, Morley S, Neely S, Lei H, Manor D & Atkinson J (2002) Expression and refolding of recombinant human α-tocopherol transfer protein capable of specific α-tocopherol binding. *Protein Expr Purif.* **24**, 395-403.

35. Smith VR & Walker JE (2003) Purification and folding of recombinant bovine oxoglutarate/malate carrier by immobilized metal-ion affinity chromatography. *Protein Expr Purif.* **29**, 209-216.

36. Batas B & Chaudhuri JB (1996) Protein refolding at high concentration using size-exclusion chromatography. *Biotechnol Bioeng.* **50**, 16-23.

37. Khan RH, Rao KB, Eshwari AN, Totey SM & Panda AK (1998) Solubilization of recombinant ovine growth hormone with retention of native-like secondary structure and its refolding from the inclusion bodies of *Escherichia coli*. *Biotechnol Prog.* **14**, 722-728.

38. St John RJ, Carpenter JF & Randolph TW (2002) High-pressure refolding of disulfide-cross-linked lysozyme aggregates: thermodynamics and optimization. *Biotechnol Prog.* **18**, 565-571.

39. Perrett S & Zhou JM (2002) Expanding the pressure technique: insights into protein folding from combined use of pressure and chemical denaturants. *Biochim Biophys Acta.* **1595**, 210-223.

40. Lefebvre BG, Gage MJ & Robinson AS (2004) Maximizing recovery of native protein from aggregates by optimizing pressure treatment. *Biotechnol Prog.* **20**, 623-629.

41. Hofmann A, Tai M, Wong W & Glabe CG (1995) A sparse-matrix screen to establish initial conditions for protein renaturation. *Anal Biochem.* **230** , 8-15.

42. Chen GQ & Gouaux E (1997) Overexpression of a glutamate receptor (GluR2) ligand binding domain in *Escherichia coli*: application of a novel protein folding screen. *Proc Natl Acad Sci USA.* **94**, 13431-13436.

43. Armstrong N, de Lencastre A & Gouaux E (1999) A new protein folding screen: application to the ligand binding domains of a glutamate and kainate receptor and to lysozyme and carbonic anhydrase. *Protein Sci.* **8**, 1475-1483.

44. Klein C, Hesse F, Dehner A, Engh RA, Schwaiger M & Hansen S (2004) In vitro folding and characterization of the p53 DNA binding domain. *Biol Chem.* **385**, 95-102.

45. Tobbell DA, Middleton BJ, Raines S, Needham MR, Taylor IW, Beveridge JY & Abbott WM (2002) Identification of *in vitro* folding conditions for procathepsin S and cathepsin S using fractional factorial screens. *Protein Expr Purif.* **24**, 242-254.

46. Builder S, Hart R, Lester P, & Reifsnyder D (1997) Refolding of misfolded insulin-like growth factor I. U.S.A. Patent number 5663304.

47. Kahmann JD, Koruth R & Day AJ (1997) Method for quantitative refolding of the link module from human TSG-6. *Protein Expr Purif.* **9**, 315-318.

48. Liang S, Shu Q, Wang X & Zong X (1999) Oxidative folding of reduced and denatured huwentoxin-I. *J Protein Chem.* **18**, 619-625.

49. Ahmed AK, Schaffer SW & Wetlaufer DB (1975) Nonenzymic reactivation of reduced bovine pancreatic ribonuclease by air oxidation and by glutathione oxidoreduction buffers. *J Biol Chem.* **250**, 8477-8482.

50. Woycechowsky KJ, Wittrup KD & Raines RT (1999) A small-molecule catalyst of protein folding *in vitro* and i*n vivo*. *Chem Biol.* **6**, 871-879.

51. Gough JD, Williams RH, Donofrio AE & Lees WJ (2002) Folding disulfide-containing proteins faster with an aromatic thiol. *J Am Chem Soc.* **124**, 3885-3892.

52. Kersteen EA & Raines RT (2003) Catalysis of protein folding by protein disulfide isomerase and small-molecule mimics. *Antioxid Redox Signal.* **5**, 413-424.

53. Wedemeyer WJ, Welker E, Narayan M & Scheraga HA (2000) Disulfide bonds and protein folding. *Biochemistry.* **39**, 4207-4216.

54. Fernandes PA & Ramos MJ (2004) Theoretical insights into the mechanism for thiol/disulfide exchange. *Chem-Eur J.* **10**, 257-266.

55. Gilbert HF (1990) Molecular and cellular aspects of thiol disulfide exchange. *Adv Enzymol Relat Areas Mol Biol.* **63**, 69-172.

56. DeCollo TV & Lees WJ (2001) Effects of aromatic thiols on thiol-disulfide interchange reactions that occur during protein folding. *J Org Chem.* **66**, 4244-4249.

57. Hiniker A & Bardwell JC (2003) Disulfide bond isomerization in prokaryotes. *Biochemistry.* **42**, 1179-1185.

58. Schwaller M, Wilkinson B & Gilbert HF (2003) Reduction-reoxidation cycles contribute to catalysis of disulfide isomerization by protein-disulfide isomerase. *J Biol Chem.* **278**, 7154-7159.

59. Fischer B, Sumner I & Goodenough P (1993) Isolation, renaturation, and formation of disulfide bonds of eukaryotic proteins expressed in Escherichia coli as inclusion bodies. *Biotechnol Bioeng.* **41**, 3-13.

60. Bernardez Clark E (1998) Refolding of recombinant proteins. *Curr Opin Biotechnol.* **9**, 157-163.

61. Bernardez Clark E, Schwarz E & Rudolph R (1999) Inhibition of aggregation side reactions during *in vitro* protein folding. *Methods Enzymol.* **309**, 217-236.

62. Misawa S & Kumagai I (1999) Refolding of therapeutic proteins produced in *Escherichia coli* as inclusion bodies. *Biopolymers.* **51**, 297-307.

63. Middelberg AP (2002) Preparative protein refolding. *Trends Biotechnol.* **20**, 437-443.

64. Jones DB, Hutchinson MH & Middelberg APJ (2004) Screening protein refolding using surface plasmon resonance. *Proteomics.* **4**, 1007-1013.

65. Heiring C & Muller YA (2001) Folding screening assayed by proteolysis: application to various cystine deletion mutants of vascular endothelial growth factor. *Protein Eng.* **14**, 183-188.

66. Scheich C, Niesen FH, Seckler R & Bussow K (2004) An automated *in vitro* protein folding screen applied to a human dynactin subunit. *Protein Sci.* **13**, 370-380.

67. DeLucas LJ, Bray TL, Nagy L, McCombs D, Chernov N, Hamrick D, Cosenza L, Belgovskiy A, Stoops B & Chait A (2003) Efficient protein crystallization. *J Struct Biol.* **142**, 188-206.

68. Sedzik J (1995) Regression analysis of factorial designed trials. A logical approach to protein crystallisation. *BBA-Protein Struc Mol Enzym.* **1251**, 177-185.

69. Xie L, Hall D, Eiteman MA & Altman E (2003) Optimization of recombinant aminolevulinate synthase production in *Escherichia coli* using factorial design. *Appl Microbiol Biotechnol.* **63**, 267-273.

70. Damaso MC, Almeida MS, Kurtenbach E, Martins OB, Pereira N, Jr, Andrade CM & Albano RM (2003) Optimized expression of a thermostable xylanase from *Thermomyces lanuginosus* in *Pichia pastoris*. *Appl Environ Microbiol.* **69**, 6064-6072.

71. Bouvier A, Chapline J, Boerner R, Jeyarajah S, Cook S, Acharya PS, Henderson I, Schrimsher JL & Shepard SR (2003) Identifying and modulating disulfide formation in the biopharmaceutical production of a recombinant protein vaccine candidate. *J Biotechnol.* **103**, 257-271.

72. Andrade VS, Neto BB, Souza W & Campos-Takaki GM (2000) A factorial design analysis of chitin production by *Cunninghamella elegans*. *Can J Microbiol.* **46**, 1042-1045.

73. Shukla AA, Sorge L, Boldman J & Waugh S (2001) Process characterization for metal-affinity chromatography of an Fc fusion protein: a design-of-experiments approach. *Biotechnol Appl Biochem.* **34**, 71-80.

74. Kalil SJ, Suzan R, Maugeri F & Rodrigues MI (2001) Optimization of inulinase production by *Kluyveromyces marxianus* using factorial design. *Appl Biochem Biotechnol.* **94**, 257-264.

75. Chun C, Heineken K, Szeto D, Ryll T, Chamow S & Chung JD (2003) Application of factorial design to accelerate identification of CHO growth factor requirements. *Biotechnol Prog.* **19**, 52-57.

76. Wildsmith SE, Archer GE, Winkley AJ, Lane PW & Bugelski PJ (2001) Maximization of signal derived from cDNA microarrays. *Biotechniques.* **30**, 202-206.

77. Box GEP, Hunter WG & Hunter JS (1978) Statistics for experimenters. An introduction to design, data analysis, and model building. John Wiley & Sons Inc, New York.

78. Studier FW & Moffatt BA (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J Mol Biol.* **189**, 113-130.

79. Wu CFJ & Hamada M (2000) Experiments. Planning, analysis, and parameter design optimization. John Wiley & Sons Inc, New York.

80. Goldenberg DP (1997) Analysis of protein conformation by gel electrophoresis. In *Protein Structure: a practical approach.* (Creighton TE ed.), pp. 187-218, Oxford : IRL Press at Oxford University Press, Oxford.

81. Fischer B, Perry B, Sumner I & Goodenough P (1992) A novel sequential procedure to enhance the renaturation of recombinant protein from *Escherichia coli* inclusion bodies. *Protein Eng.* **5**, 593-596.

82. Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL & Sykes BD (1995) $^{1}H$, $^{13}C$ and $^{15}N$ chemical shift referencing in biomolecular NMR. *J Biomol NMR.* **6**, 135-140.

83. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J & Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* **6**, 277-293.

84. Boucher, W (2002) AZARA suite for NMR data (version 2.7) Department of Biochemistry, University of Cambridge.

85. Richards F & Wyckoff H (1971) The Enzymes. Academic Press, New York.

86. Alderton G, Ward W & Febold H (1945) Isolation of lysozyme from egg white. *J Biol Chem.* **157**, 43-58.

87. Maachupalli-Reddy J, Kelley BD & De Bernardez CE (1997) Effect of inclusion body contaminants on the oxidative renaturation of hen egg white lysozyme. *Biotechnol Prog.* **13**, 144-150.

88. Tran-Moseman A, Schauer N & De Bernardez CE (1999) Renaturation of *Escherichia coli*-derived recombinant human macrophage colony-stimulating factor. *Protein Expr Purif.* **16**, 181-189.

89. Futami J, Tsushima Y, Tada H, Seno M & Yamada H (2000) Convenient and efficient *in vitro* folding of disulfide-containing globular protein from crude bacterial inclusion bodies. *J Biochem (Tokyo).* **127**, 435-441.

90. Ouellette T, Destrau S, Ouellette T, Zhu J, Roach JM, Coffman JD, Hecht T, Lynch JE & Giardina SL (2003) Production and purification of refolded recombinant human IL-7 from inclusion bodies. *Protein Expr Purif.* **30**, 156-166.

91. Speed MA, Wang DI & King J (1996) Specific aggregation of partially folded polypeptide chains: the molecular basis of inclusion body composition. *Nat Biotechnol.* **14**, 1283-1287.

92. Wuthrich K (1986) NMR of protein and nucleic acids. John Wiley-Interscience,

93. Halfmann G, Brailly H, Bernadac A, Montero-Julian FA, Lazdunski C & Baty D (1993) Targeting of interleukin-2 to the periplasm of *Escherichia coli*. *J Gen Microbiol.* **139**, 2465-2473.

94. McCallus DE, Ugen KE, Sato AI, Williams WV & Weiner DB (1992) Construction of a recombinant bacterial human CD4 expression system producing a bioactive CD4 molecule. *Viral Immunol.* **5**, 163-172.

95. Kurokawa Y, Yanagi H & Yura T (2001) Overproduction of bacterial protein disulfide isomerase (DsbC) and its modulator (DsbD) markedly enhances periplasmic production of human nerve growth factor in *Escherichia coli*. *J Biol Chem.* **276**, 14393-14399.

96. Bessette PH, Aslund F, Beckwith J & Georgiou G (1999) Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *Proc Natl Acad Sci USA.* **96**, 13703-13708.

97. Derman AI, Prinz WA, Belin D & Beckwith J (1993) Mutations that allow disulfide bond formation in the cytoplasm of *Escherichia coli*. *Science.* **262**, 1744-1747.

98. LaVallie ER, DiBlasio EA, Kovacic S, Grant KL, Schendel PF & McCoy JM (1993) A thioredoxin gene fusion expression system that circumvents inclusion body formation in *Escherichia coli* cytoplasm. *Bio-Technology.* **11**, 187-193.

99. Day AJ, Aplin RT & Willis AC (1996) Overexpression, purification, and refolding of link module from human TSG-6 in *Escherichia coli*: effect of temperature, media, and mutagenesis on lysine misincorporation at arginine AGA codons. *Protein Expr Purif.* **8**, 1-16.

100. Zhang JG, Moritz RL, Reid GE, Ward LD & Simpson RJ (1992) Purification and characterization of a recombinant murine interleukin-6. Isolation of N- and C-terminally truncated forms. *Eur J Biochem.* **207**, 903-913.

101. Han B, Hall FL & Nimni ME (1997) Refolding of a recombinant collagen-targeted TGF-α2 fusion protein expressed in *Escherichia coli*. *Protein Expr Purif.* **11**, 169-178.

102. Banerji S, Day AJ, Kahmann JD & Jackson DG (1998) Characterization of a functional hyaluronan-binding domain from the human CD44 molecule expressed in *Escherichia coli*. *Protein Expr Purif.* **14**, 371-381.

103. Boyle DM, McKinnie RE, Joy WD, Violand BN, McLaughlin JK & Landis BH (1999) Evaluation of refolding conditions for a recombinant human interleukin-3 variant (daniplestim). *Biotechnol Appl Biochem.* **30**, 163-170.

104. Ahn JH, Lee YP & Rhee JS (1997) Investigation of refolding condition for *Pseudomonas fluorescens* lipase by response surface methodology. *J Biotechnol.* **54**, 151-160.

105. (1999) Hydrophobic Interaction Chromatography. Principles and Methods (Amersham, Pharmacia Biotech). Uppsala Sweden.

106. Titani K, Sasagawa T, Resing K & Walsh KA (1982) A simple and rapid purification of commercial trypsin and chymotrypsin by reverse-phase high-performance liquid chromatography. *Anal Biochem.* **123**, 408-412.

107. Bischoff R, Clesse D, Whitechurch O, Lepage P & Roitsch C (1989) Isolation of recombinant hirudin by preparative high-performance liquid chromatography. *J Chromatogr.* **476**, 245-255.

108. Urdal DL, Mochizuki D, Conlon PJ, March CJ, Remerowski ML, Eisenman J, Ramthun C & Gillis S (1984) Lymphokine purification by reversed-phase high-performance liquid chromatography. *J Chromatogr.* **296**, 171-179.

109. van den Berg B, Chung EW, Robinson CV, Mateo PL & Dobson CM (1999) The oxidative refolding of hen lysozyme and its catalysis by protein disulfide isomerase. *EMBO J.* **18**, 4794-4803.

110. Raftery MJ, Collinson L & Geczy CL (1999) Overexpression, oxidative refolding, and zinc binding of recombinant forms of the murine S100 protein MRP14 (S100A9). *Protein Expr Purif.* **15**, 228-235.

# Chapter Three

# Improvement of solubility and *in vitro* oxidative folding yield of the N-terminal domain of human CD5 by site-directed mutagenesis


## Abstract

The poor solubility of the N-terminal domain of human CD5 (*h*CD5d1) expressed in *Escherichia coli* makes it unsuitable for structural studies. Replacing hydrophobic residues for polar amino acids to improve solubility has been successful in other cases and this approach has been adopted here. In order to find target residues for substitution, the amino acid sequence of *h*CD5d1 was compared with orthologue sequences and with that of the scavenger receptor cysteine-rich domain of Mac-2 binding protein, the only homologous structure known at that time. Eight single mutations were introduced in positions that were occupied by non-polar residues in *h*CD5d1 and by polar residues in at least one of the other sequences. Similarly to the wild type *h*CD5d1, the eight single mutants expressed in high yields as inclusion bodies. However, the *in vitro* folding yield of four of the mutants was higher than that of the wild type. The best two variants were combined in a double substitution mutant: V88D/V97K *h*CD5d1. The folding yield of this double mutant is four times higher than the wild type. Furthermore this new variant could be concentrated to more than 1.0 mM, making V88D/V97K *h*CD5d1 appropriate for structural studies by multidimensional NMR.

Chapter abbreviations and acronyms

**3D** three-dimensional

**ε** molar extinction coefficient

**ρ** Spearman correlation coefficient

**ANOVA** analysis of variance

**GdnHCl** guanidine hydrochloride

***h*CD5d1** N-terminal SRCR domain of CD5

**HPLC** high performance liquid chromatography

**HSQC** heteronuclear single quantum coherence

**IPTG** isopropyl-β-D-thiogalactopyranoside

**M2BP** Mac-2 binding protein

**NMR** nuclear magnetic resonance

**PDB** Protein Data Bank

**RPC** reversed phase chromatography

**SEC** size-exclusion chromatography

**SRCR** scavenger receptor cysteine-rich

**WT** wild type

## Introduction

The attainment of high protein concentrations is paramount in structural biology. Crystallisation trials require protein concentrations ranging from around 2 mg/ml to 20 mg/ml. Multidimensional NMR experiments typically require protein concentrations of at least 0.5 mM and preferentially greater that 1 mM, in order to detect signals in some of the intrinsically insensitive experiments required for resonance assignment. Particularly for NMR, a considerable amount of time is spent in optimising the solution composition to maximise protein solubility. Frequently, even at optimal solubility conditions, the required protein concentration cannot be achieved, and can become necessary to make modifications to the protein sequence in order to enhance solubility.

Amino acid substitution has proved to be a powerful technique for manipulating the stability and solubility of recombinant proteins. The technique has been used mainly to reduce the formation of inclusion bodies, but also to enhance the stability and/or solubility of purified proteins. The improvement of the yield of soluble protein *in vivo* using random mutagenesis and an appropriate selection criterion has become a popular technique in structural genomics initiatives. The solubility of the mutant proteins is assessed *in vivo* by means of a gene fusion reporter system, such as the fluorescence emitted by the green fluorescent protein[1], the resistance to antibiotic conferred by the choramphenicol acetyltransferase[2] or the development of blue colour as a result of the hydrolysis of X-Gal (5-bromo-4-chloro-3-indolyl-β-galactopyranoside) by β-galactosidase[3]. However, to set up a successful 'directed evolution' protocol often

requires a considerable amount of time and effort and may necessitate some degree of automation to attain a good mutation rate.

In small scale protein expression projects, the rational design of mutations may be a more efficient strategy to follow. A very useful approach has been to substitute solvent-exposed hydrophobic residues for hydrophilic amino acids. This approach has been successful in enhancing the yield of correctly folded protein *in vivo*[4-7] and in improving the solubility of purified proteins[8-12]. Even in the absence of structural information to evaluate solvent accessibility, groups have succeeded in using this approach by exhaustively replacing hydrophobic residues[6]. The construction of large sets of mutants has become a relatively straightforward task, because of the availability of efficient commercial site-directed mutagenesis kits that do not require the DNA of interest to be subcloned in specific vectors or necessitate the isolation of single-stranded DNA. Nevertheless, to maximise the efficiency of this approach it is sensible to more precisely target the mutation sites by comparing the target protein amino acid sequence with orthologue sequences. The analysis of a multiple sequence alignment highlights conserved hydrophobic residues, these are likely to be involved in the hydrophobic packing of the core of the protein and therefore are not appropriate for mutation to improve solubility. On the other hand, residues that are hydrophobic in the target sequence but hydrophilic in an orthologue are good candidates for mutation, as they have a high likelihood of preserving the fold and enhancing solubility. This approach has been successfully applied for the catalytic domain of β-1,4-galactosyltransferase 1[10] and the adhesion domain of CD58[8]. If available, the analysis of the 3D structure of a homologous fold can provide additional information about any particular residue's likely solvent accessibility and help identify residues with an important structural role that should not be altered[7,8,11].

In the previous Chapter, I reported the expression, purification and optimisation of the *in vitro* folding conditions for a construct of the N-terminal scavenger receptor cysteine-rich domain of human leukocyte receptor CD5 (*h*CD5d1). However, the success of this work was limited, because even the optimised folding yields were low overall and it was not possible to concentrate the folded *h*CD5d1 to higher than 0.3 mM, whereas a concentration around 1 mM is preferred to attempt structure determination. In this Chapter, I describe the improvement of *h*CD5d1 solubility and *in vitro* folding yield

using a directed mutagenesis approach. Some observations on the *in vitro* folding kinetics of the enhanced solubility mutant are also presented.

## Experimental procedures[1]

*Design of hCD5d1 mutants*

In order to identify target residues for mutation, the amino acid sequence of *h*CD5d1 was aligned with the orthologue sequences for other mammals using ClustalW[13]. Amino acid sequences were obtained from the SwissProt and Entrez protein sequence databases: mouse CD5 (SP, P13379), rat CD5 (SP, P51882), cow CD5 (SP, P19238) and sheep CD5 (EP, I47074). The CD5d1 sequences were manually aligned with the sequence for the scavenger receptor cysteine-rich (SRCR) domain of Mac-2 binding protein[14] (M2BP SRCR; PDB code 1by2), the only homologous structure known at that time. Residue by residue side chain solvent accessibilities for the M2BP structure were calculated using the program NACCESS[15].

*Site-directed mutagenesis*

Mutants of the construct comprising residues 25 to 134 of *h*CD5 (*h*CD5d1) were made using the QuickChange Site-Directed Mutagenesis kit (Stratagene), following the manufacturer's instructions. Oligonucleotides were purchased from MWG-Biotech AG (see Appendix B for sequences). *E. coli* XL1-Blue cells were transformed with the mutagenesis products and plasmid DNA was isolated from 10 ml of an overnight cell culture using the Qiaprep spin Miniprep kit (Qiagen). The presence of the mutations was verified by automated DNA sequencing using dRhodamine dye-terminator chemistry (Applied Biosciences). The sequencing reaction products were run on an Abi Prism 377 sequencer at the Ludwig Institute of Cancer Research by Dr Athena Nikitopoulou.

[1] Full details of the protocols are provided in Appendix A.

*Overexpression and purification of hCD5d1 mutants*

The methodology for the production of mutant *h*CD5d1 protein products was the same as that described for the wild type protein in Chapter II. Briefly, each of two 500 ml of LB medium with 50μg/ml carbenicillin in a 2-litre baffled flask was inoculated with 1 ml of an overnight culture grown at room temperature with shaking. The cultures were incubated at 37 °C with vigorous shaking until the $OD_{600}$ reached 1.0. Protein expression was then induced by addition of IPTG to a final concentration of 1 mM. Cells were harvested 4-12 hours after induction by centrifugation, resuspended, and lysed using a French pressure cell. The inclusion bodies were isolated by centrifugation, washed and solubilised in 6 M guanidine hydrochloride. The overexpression product was purified by immobilised metal ion affinity chromatography in denaturing conditions and elution was achieved by reducing the buffer pH to 4.5. The eluate was reduced with 0.1 M DTT. The *h*CD5d1 monomer was finally isolated by size-exclusion chromatography (SEC) under denaturing conditions.

For the production of $^{15}N$-labelled samples, 1 ml of an 8-hour culture in LB with 50μg/ml carbenicillin was used to inoculate 100 ml of unlabelled M9 minimal medium with supplements and antibiotic. The cultures were incubated overnight at 30 °C with shaking. Cells were harvested by centrifugation and resuspended in fresh minimal medium. A volume equivalent to 10 ml of the overnight culture was used to inoculate 500 ml of $^{15}N$-labelled minimal medium with 50 μg/ml carbenicillin. From this point, the same methodology used for the overexpression and purification of un-labelled samples was followed.

*In vitro folding of mutants*

The monomer-containing fractions (fractions 5 and 6; Figure III.3B) from SEC were concentrated to more than 10 mg/ml by centrifugal ultrafiltration (Vivaspin 20, Vivascience). Immediately after concentration, folding was performed at 4°C by rapid dilution of the concentrated protein sample into a 0.1 M Tris-HCl buffer pH 9 with 0.6 M arginine, 20% glycerol, 5 mM cysteine and 0.5 mM cystine (*i.e.* the best buffer found in the study of the *in vitro* folding of the wild type protein in Chapter II). Samples were incubated for 24 hours with gentle stirring and then clarified by centrifugation and dialysed against a 50 mM phosphate buffer pH 6.0 with 0.2 M NaCl, 1 mM EDTA and 0.05% (w/v) sodium azide.

*Kinetic observations*

*In vitro* folding of V88D/V97K *h*CD5d1 was initiated as described above and 5 ml samples were taken at the desired time points. The disulphide bond formation reaction in the sample was immediately quenched with 50% (v/v) acetic acid, which lowered the pH below 2.0 (the oxidation reaction is no longer favourable under these conditions).

*Reversed-phase HPLC*

Acid-quenched and dialysed folding samples of V88D/V97K *h*CD5d1 were analysed using a Sephasil Protein C4 (300 Å, 5μm, 4.6 mm x 100 mm) RP-HPLC column connected to an Äkta Purifier chromatography system (Amersham Biosciences). Detection was performed by measuring absorbance at 220 and 280 nm. Solvent A was 5% (v/v) acetonitrile with 0.1% (v/v) trifluoroacetic acid (TFA) and solvent B was 95% acetonitrile 0.1% TFA.

For the *in vitro* folding screen of *h*CD5d1 mutants, samples were eluted using a linear gradient from 0% to 45% B in 15 minutes with a flow rate of 2 ml/min. For the kinetic observations of V88D/V97K *h*CD5d1 folding a linear gradient from 28% to 40% B was run in 8 minutes with a flow rate of 0.5 ml/min.

*Estimation of protein concentration*

Protein concentration was estimated by measuring absorbance at 280 nm using the theoretical molar extinction coefficient calculated with the Gill and von Hippel equation[16]. This is 27880 $M^{-1}cm^{-1}$ for the majority of the constructs, except for W28R for which the theoretical molar extinction coefficient is 22180 $M^{-1}cm^{-1}$.

*NMR spectroscopy*

The $^{1}H$-$^{15}N$ HSQC spectra were acquired at 298 K on a Varian UnityPlus spectrometer operating at a nominal $^{1}H$ frequency of 600 MHz. The spectrum of V88D/V97K *h*CD5d1 was recorded with 1024 complex points and a spectral width of 5 kHz in the $^{1}H$ dimension and 256 complex points and a spectral width of 1.7 kHz in the $^{15}N$ dimension. The data for the *h*CD5d1 construct expressed in *Pichia pastoris* was provided by Dr Mark Pfuhl (Leicester University). The spectrum was recorded with 1600 complex points and a spectral width of 5.5 kHz in the $^{1}H$ dimension and 256 complex points and a spectral width of 2.0 kHz in the $^{15}N$ dimension. Chemical shifts were indirectly referenced to sodium 2,2-dimethyl-2-silane-pentane-5-sulphonate (DSS) using absolute frequency ratios for the $^{1}H$ signals[17]. All data were processed using NMRpipe/NMRDraw[18]. Figure III.7 was made with the Plot2 program of the AZARA suite[19].

## Results

*Design of hCD5d1 mutants*

The *h*CD5d1 amino acid sequence was aligned with the sequences from CD5 orthologues in order to find positions that were occupied by non-polar amino acids in the human protein and by polar amino acids in any of the other sequences. There are nine positions that fulfil this criterion: L26, W28, M58, W71, V88, L96, V97, M124 and G129 (Figure III.1). The amino acid sequence of the M2BP SRCR domain was also aligned with the CD5d1 sequences in order to derive some structural information for the mutation candidates. The SRCR domain of M2BP shares 18% identity and 28% similarity to *h*CD5d1. This is a low homology level and structural inferences from M2BP SRCR were taken with due caution. On the basis of this analysis the residues selected for mutation were L26, W28, V88, V97, M124 and G129. Residues M58 and L96 were not considered for mutation because the equivalent residues are buried in the M2BP SRCR structure. Gly129 is also buried in the M2BP SRCR structure. Nevertheless, this residue was targeted for mutation because the accessibility of residues with larger side chains are not well predicted from the values reported for Gly. Trp71 was not chosen for mutation because it lies in an insertion not present in the M2BP SRCR the structure of which may have a particular biological significance.

A total of eight single-site mutants of *h*CD5d1 were constructed: L26K, W28R, V88D, V88N, V97K, M124N, M124Q and G129S. Of these L26K was the only mutation to an amino acid type not present in an orthologue sequence. This N-terminal residue was not expected to be part of the globular fold and a less conservative approach was taken to its mutation. It was reasoned that a change of Leu to Lys as opposed to the Ser found in some of the other sequences, would further enhance solubility of the protein.

After evaluating the *in vitro* folding performance of the single mutants as described below, the mutations with the best performance, were combined in a double mutant construct called V88D/V97K *h*CD5d1.

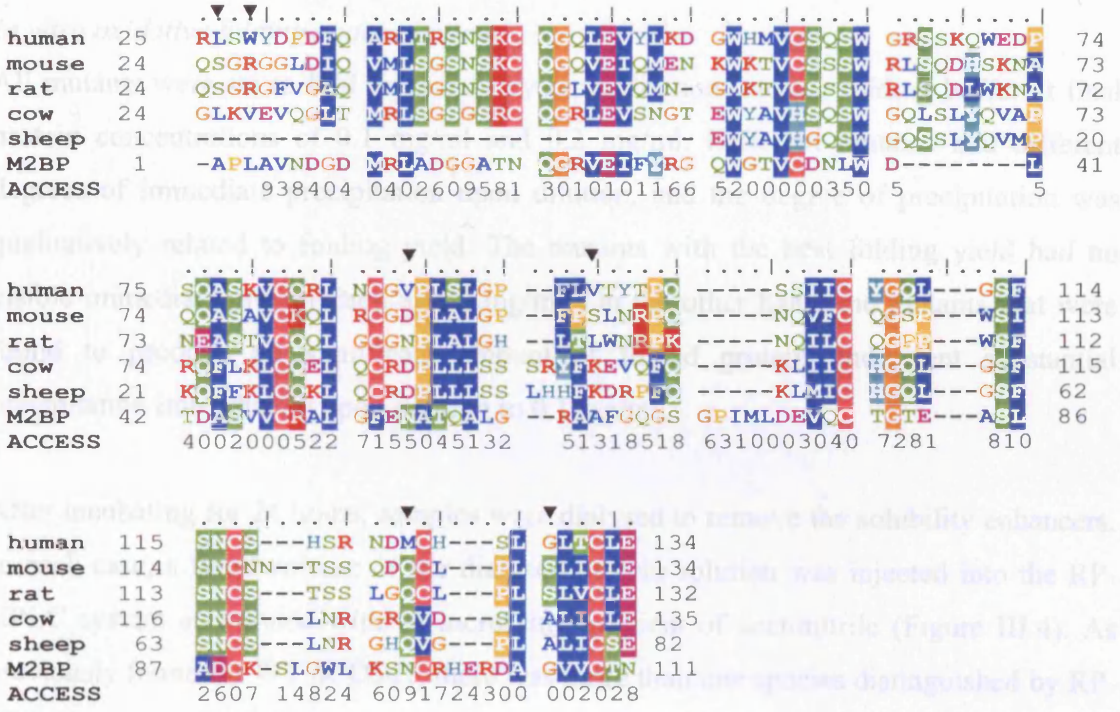


**Figure III.1** Alignment of the amino acid sequences of the N-terminal domain of CD5 (CD5d1) of different species and the SRCR domain of Mac-2 binding protein (M2BP).
Arrows indicate residues selected for mutation in *h*CD5d1.
The accessibility values correspond to the side chain solvent accessibility as computed by NACCESS for the SRCR domain of M2BP. The scale runs from non-accessible(0) to very accessible (9).
The sequence of the sheep CD5d1 appears truncated in the database.

*Expression and purification of hCD5d1 mutants*

The mutant constructs, each bearing an N-terminal hexa-histidine tag, were expressed in BL21 (DE3) cells using the IPTG-inducible T7 promoter system. Similarly to the WT protein, all mutants expressed in high yields as inclusion bodies (Figure III.2). Each mutant protein was solubilised using 6 M GdnHCl, purified by immobilised metal ion affinity chromatography, cysteines were reduced and the monomeric protein chains were isolated by SEC. As far as expression and purification are concerned, there were no significant differences among the mutants. Figure III.3 shows the results of a typical protein purification procedure.

*In vitro oxidative folding mutant screen*

All mutants were set to fold separately by rapid dilution into the folding buffer at final protein concentrations of 0.1 mg/ml and 0.2 mg/ml. Different mutants had different degrees of immediate precipitation upon dilution, and the degree of precipitation was qualitatively related to folding yield. The mutants with the best folding yield had no visible immediate precipitation at 0.2 mg/ml. On the other hand, the mutants that were found to produce no significant amount of folded protein underwent substantial precipitation immediately upon dilution to 0.1 mg/ml.

After incubating for 24 hours, samples were dialysed to remove the solubility enhancers. In each case, a fixed volume of the dialysed protein solution was injected into the RP-HPLC system and eluted with an increasing gradient of acetonitrile (Figure III.4). As previously found for WT *h*CD5d1, there was more than one species distinguished by RP-HPLC in most of the dialysed protein solutions. The native fold of a protein is expected to minimise its hydrophobicity, consequently it is commonly assumed that the lowest retention volume peak in RP chromatography corresponds to the natively folded species. The results shown in Chapter II strongly support this supposition, as the buffer conditions which maximised the area for the lowest retention volume peak in RP-HPLC experiments, were also those that yielded a high proportion of folded WT *h*CD5d1 as validated by NMR. It was assumed that this was the case for all *h*CD5d1 mutants and that the peak that eluted first was most likely to be the monomeric native folded species. The first peak for each of the mutants eluted within a narrow range of retention volumes and several clearly shown elution volumes lower than that of the WT protein,

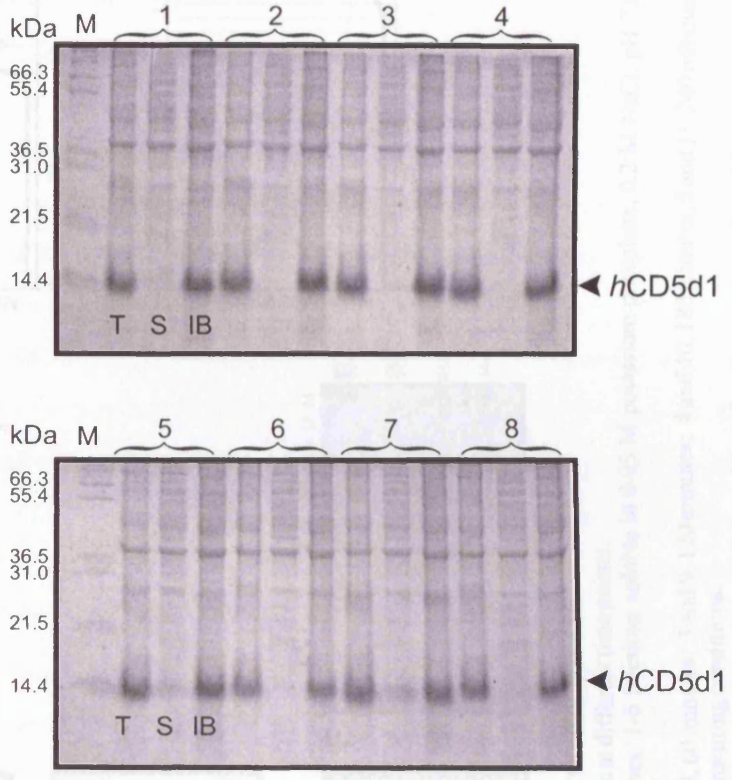


**Figure III.2** Non-reducing SDS-PAGE analysis of the expression of different single-residue mutants of *h*CD5d1 in BL21(DE3) cells.
Lane M shows molecular weight markers (Mark 12, Invitrogen). There are three lanes for each mutant always in the same order: **T** total protein; **S** fraction of protein soluble in 50 mM phosphate buffer pH 7.0; **IB** fraction soluble in 8 M urea but insoluble in 50 mM phosphate buffer pH 7.0.
Groups of three lanes: 1, L26K; 2, W28R; 3, V88D; 4, V88K; 5, V97K: 6, M124N; 7, M124Q; 8, G129S.

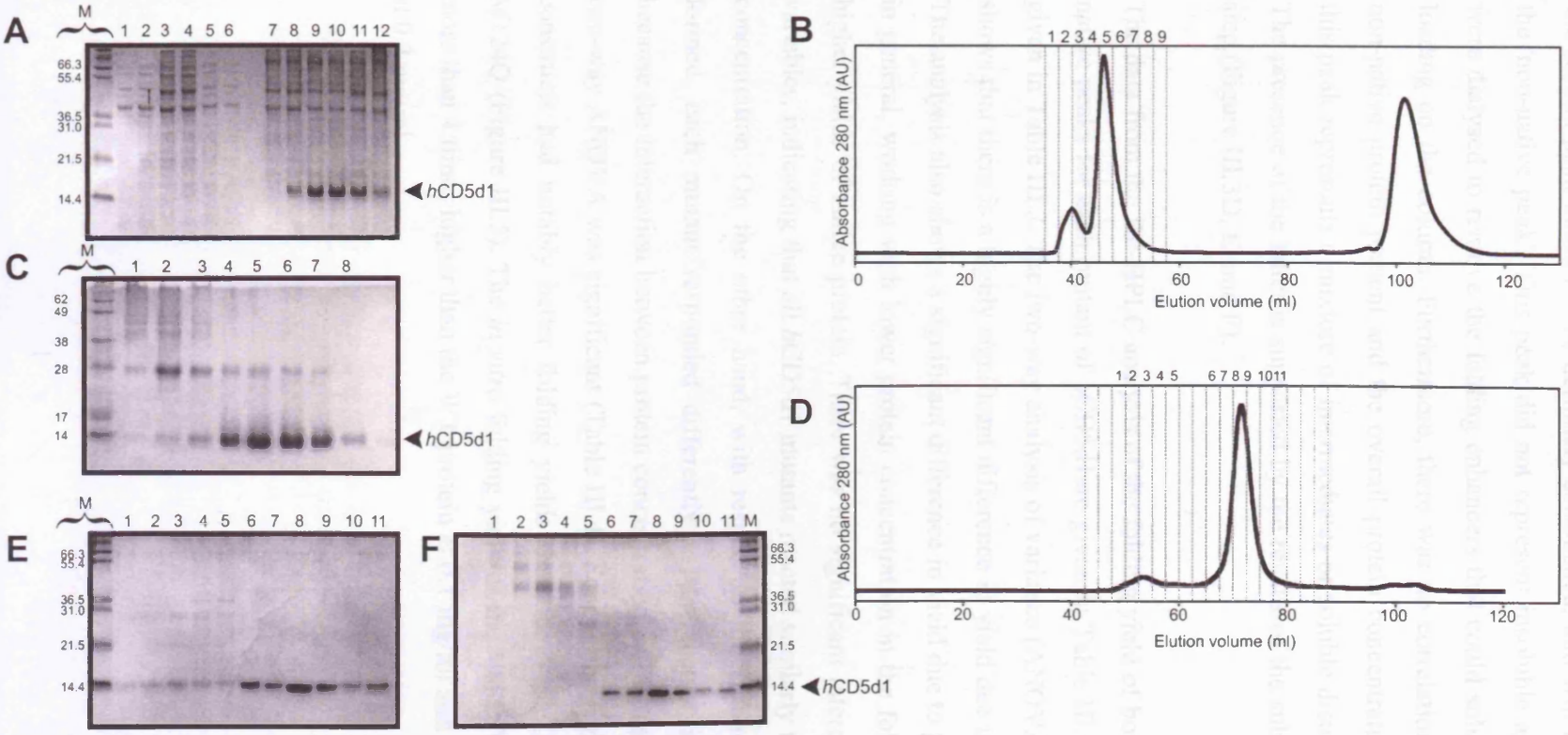


**Figure III.3** Expression and purification of V88D/V97K *h*CD5d1.
(A) Reducing SDS-PAGE analysis of expression at different time points.
Lane M: molecular weight markers (kDa). Lanes: 1-6 fractions soluble in 0.05 M potassium phosphate, 0.2 M NaCl, pH 7.0. Lanes 7-12: fractions soluble in 8 M urea, 0.1 M potassium phosphate, 0.01 M Tris, pH 8.0.
Lanes: 1 and 7, 0 minutes after induction; 2 and 8, 60 minutes; 3 and 9, 120 minutes; 4 and 10, 180 minutes; 5 and 11, 240 minutes; 6 and 12, 960 minutes after induction.
(B) Size exclusion chromatography (SEC) in denaturing conditions.
(C) Non-reducing SDS-PAGE analysis of fractions from denaturing SEC. Lane M: molecular weight markers (kDa). Other lane numbers correspond to the fraction number in B.
(D) Size exclusion chromatography of refolded protein.
(E) Reducing SDS-PAGE analysis of fractions from SEC of refolded protein. Lane M: molecular weight markers (kDa). Other lane numbers correspond to the fraction number in D. The gel shows that the high molecular weight fractions (lanes 1-5) correspond to multiples of the monomeric protein.

suggesting that they had the desired property of increased hydrophilicity. The identity of the second peak could not be definitely established, and will be referred to hereafter as the 'non-native peak'. This peak did not represent insoluble aggregates, as the samples were dialysed to remove the folding enhancers that could solubilise aggregates prior to loading on the column. Furthermore, there was no correlation between the amount of non-native protein present and the overall protein concentration (Table III.4). Perhaps this peak represents a mixture of intermediates or soluble disulphide-bonded oligomers. The presence of the latter is supported by the results of the subsequent SEC purification step (Figure III.3D, E, and F).

The data from the RP-HPLC analysis of the folding yield of both the native and the non-native peaks for each mutant of *h*CD5d1are given in Table III.1. The average yields are given in Table III.2. The two-way analysis of variance (ANOVA) of the native peak data shows that there is a highly significant difference in yield due to mutations (Table III.3). The analysis also shows a significant difference in yield due to protein concentration, *i.e.* in general, working with lower protein concentration in the folding reaction resulted in higher yields of native protein. There was no significant interaction between these two variables, indicating that all *h*CD5d1 mutants reacted similarly to the variation in protein concentration. On the other hand, with respect to the amount of non-native species formed, each mutant responded differently to the change in protein concentration, because the interaction between protein concentration and identity of the construct in the two-way ANOVA was significant (Table III.4). Four of the eight single mutant *h*CD5d1 constructs had notably better folding yields than the WT: L26K, V88D, V97K and M124Q (Figure III.5). The *in vitro* folding yield of the V88D/V97K double mutant was more than 4 times higher than the WT protein at 0.1 mg/ml and more than 5 times higher at 0.2 mg/ml.

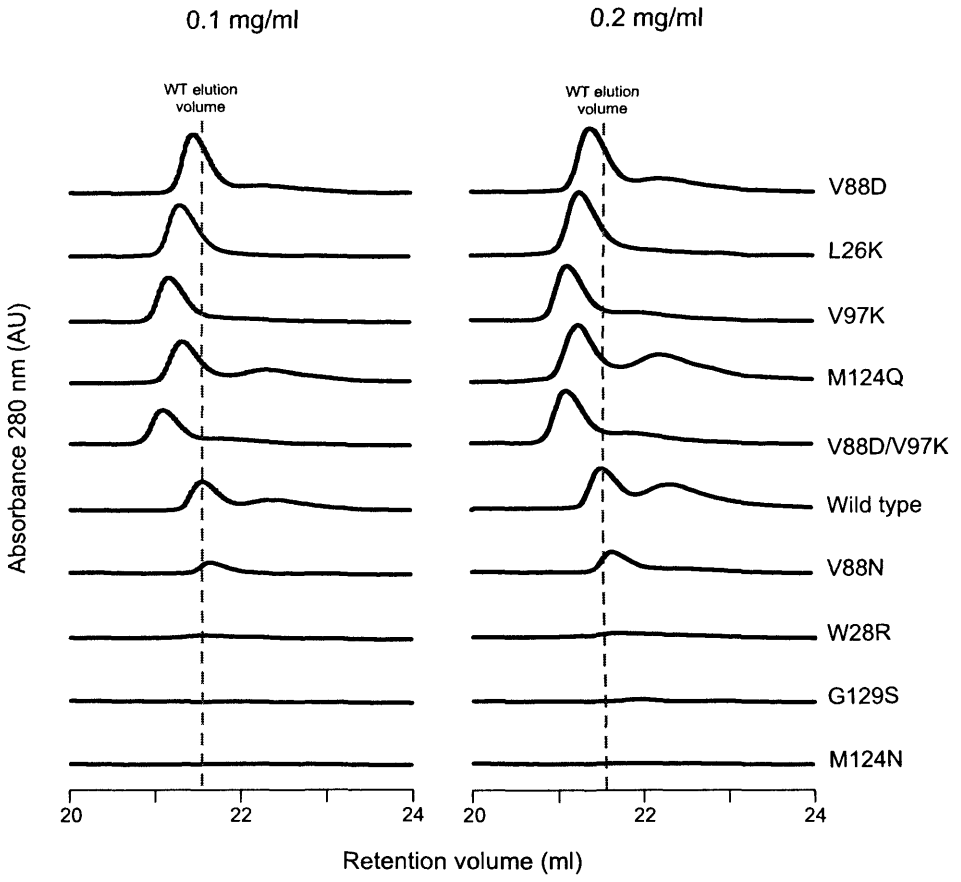


**Figure III.4** RP-HPLC elution profiles of the *h*CD5d1 mutants after *in vitro* folding, at final protein concentrations of 0.1 mg/ml and 0.2 mg/ml. The samples for V97K and V88D/V97K were diluted 1.5 and 2-fold respectively prior to injection. Evidence indicates that the peak that elutes first may correspond to the folded species and the second to an un-identified species (herein refered to as the non-native peak).

**Table III.1 Results for the *in vitro* folding analysis of *h* CD5d1 mutants (1 of 2)**

| | Construct | Replicate | Protein conc. [mg/ml] | Dialysis dilution | Injected volume [ml] | First RP-HPLC peak (native) | | | Second RP-HPLC peak (non-native) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Retention volume [ml] | Peak area | Relative yield | Retention volume [ml] | Peak area | Relative yield |
| 1 | Wild type | 1 | 0.10 | 1.29 | 2.00 | 21.43 | 8.96 | 57.13 | 22.24 | 10.71 | 68 |
| 2 | Wild type | 2 | 0.10 | 1.36 | 2.00 | 21.50 | 11.73 | 78.95 | 22.35 | 12.87 | 87 |
| 3 | Wild type | 3 | 0.10 | 1.33 | 2.00 | 21.53 | 7.77 | 51.19 | 22.38 | 6.51 | 43 |
| 4 | Wild type | 1 | 0.20 | 1.36 | 2.00 | 21.49 | 9.69 | 32.61 | 22.30 | 13.53 | 46 |
| 5 | Wild type | 2 | 0.20 | 1.36 | 2.00 | 21.52 | 11.07 | 37.25 | 22.32 | 12.88 | 43 |
| 6 | Wild type | 3 | 0.20 | 1.31 | 2.00 | 21.49 | 10.97 | 35.75 | 22.30 | 14.80 | 48 |
| 7 | L26K | 1 | 0.11 | 1.23 | 2.00 | 21.30 | 14.49 | 81.28 | 22.91 | 0.26 | 1 |
| 8 | L26K | 2 | 0.11 | 1.19 | 2.00 | 21.26 | 19.30 | 104.48 | 21.82 | 3.41 | 18 |
| 9 | L26K | 3 | 0.11 | 1.23 | 2.00 | 21.30 | 13.67 | 76.68 | - | 0.00 | 0 |
| 10 | L26K | 1 | 0.22 | 1.29 | 2.00 | 21.26 | 21.05 | 61.78 | - | 0.00 | 0 |
| 11 | L26K | 2 | 0.22 | 1.23 | 2.00 | 21.26 | 19.30 | 54.13 | 21.82 | 3.41 | 10 |
| 12 | L26K | 3 | 0.22 | 1.41 | 2.00 | 21.28 | 15.95 | 51.49 | 21.62 | 4.75 | 15 |
| 13 | W28R | 1 | 0.11 | 1.31 | 2.00 | 21.46 | 1.56 | 11.88 | - | 0.00 | 0 |
| 14 | W28R | 2 | 0.11 | 1.36 | 2.00 | 21.56 | 1.77 | 13.91 | - | 0.00 | 0 |
| 15 | W28R | 3 | 0.11 | 1.30 | 2.00 | 21.62 | 0.58 | 4.37 | 22.38 | 0.17 | 1 |
| 16 | W28R | 1 | 0.21 | 1.33 | 2.00 | 21.52 | 2.73 | 10.50 | - | 0.00 | 0 |
| 17 | W28R | 2 | 0.21 | 1.36 | 2.00 | 21.56 | 2.14 | 8.41 | - | 0.00 | 0 |
| 18 | W28R | 3 | 0.21 | 1.37 | 2.00 | 21.50 | 2.47 | 9.81 | - | 0.00 | 0 |
| 19 | V88D | 1 | 0.11 | 1.44 | 2.00 | 21.53 | 9.41 | 63.45 | - | 0.00 | 0 |
| 20 | V88D | 2 | 0.11 | 1.30 | 2.00 | 21.43 | 20.15 | 122.66 | 22.34 | 0.74 | 5 |
| 21 | V88D | 3 | 0.11 | 1.40 | 2.00 | 21.41 | 22.09 | 144.81 | 22.33 | 1.45 | 10 |
| 22 | V88D | 1 | 0.21 | 1.40 | 2.00 | 21.40 | 20.67 | 67.75 | 22.34 | 1.16 | 4 |
| 23 | V88D | 2 | 0.21 | 1.50 | 2.00 | 21.36 | 22.93 | 80.53 | 22.17 | 16.20 | 57 |
| 24 | V88D | 3 | 0.21 | 1.40 | 2.00 | 21.38 | 24.74 | 81.09 | 22.28 | 1.87 | 6 |
| 25 | V88N | 1 | 0.10 | 1.39 | 2.00 | 21.63 | 3.02 | 20.92 | - | 0.00 | 0 |
| 26 | V88N | 2 | 0.10 | 1.33 | 2.00 | 21.59 | 5.44 | 36.14 | 22.32 | 0.83 | 6 |
| 27 | V88N | 3 | 0.10 | 1.41 | 2.00 | 21.80 | 0.99 | 7.00 | - | 0.00 | 0 |
| 28 | V88N | 1 | 0.20 | 1.44 | 2.00 | 21.49 | 13.19 | 47.58 | 22.27 | 11.49 | 41 |
| 29 | V88N | 2 | 0.20 | 1.43 | 2.00 | 21.61 | 4.14 | 14.79 | - | 0.00 | 0 |
| 30 | V88N | 3 | 0.20 | 1.41 | 2.00 | 21.65 | 4.96 | 17.54 | 22.49 | 0.15 | 1 |

**Table III.1 Results for the *in vitro* folding analysis of *h* CD5d1 mutants (2 of 2)**

| | Construct | Replicate | Protein conc. [mg/ml] | Dialysis dilution | Injected volume [ml] | First peak RP-HPLC (native) | | | Second peak RP-HPLC (non-native) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Retention volume [ml] | Peak area | Relative yield | Retention volume [ml] | Peak area | Relative yield |
| 1 | V97K | 1 | 0.10 | 1.25 | 0.67 | 21.10 | 10.48 | 188.54 | 22.06 | 0.91 | 16 |
| 2 | V97K | 2 | 0.10 | 1.38 | 0.67 | 21.10 | 10.59 | 209.57 | 22.02 | 0.90 | 18 |
| 3 | V97K | 3 | 0.10 | 1.23 | 0.67 | 21.15 | 12.02 | 213.37 | 22.09 | 1.13 | 20 |
| 4 | V97K | 1 | 0.21 | 1.31 | 0.67 | 21.11 | 15.43 | 145.74 | 22.11 | 2.24 | 21 |
| 5 | V97K | 2 | 0.21 | 1.38 | 0.67 | 21.12 | 15.00 | 148.42 | 22.13 | 1.94 | 19 |
| 6 | V97K | 3 | 0.21 | 1.36 | 0.67 | 21.12 | 13.91 | 136.39 | 22.11 | 1.56 | 15 |
| 7 | M124N | 1 | 0.10 | 1.36 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 8 | M124N | 2 | 0.10 | 1.34 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 9 | M124N | 3 | 0.10 | 1.29 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 10 | M124N | 1 | 0.20 | 1.36 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 11 | M124N | 2 | 0.20 | 1.37 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 12 | M124N | 3 | 0.20 | 1.34 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 43 | M124Q | 1 | 0.10 | 1.41 | 2.00 | 21.31 | 12.07 | 83.15 | 22.31 | 7.65 | 53 |
| 44 | M124Q | 2 | 0.10 | 1.36 | 2.00 | 21.26 | 12.42 | 82.11 | 22.22 | 9.59 | 63 |
| 45 | M124Q | 3 | 0.10 | 1.37 | 2.00 | 21.25 | 14.35 | 95.86 | 22.23 | 10.72 | 72 |
| 46 | M124Q | 1 | 0.21 | 1.31 | 2.00 | 21.24 | 17.04 | 54.55 | 22.19 | 18.72 | 60 |
| 47 | M124Q | 2 | 0.21 | 1.36 | 2.00 | 21.25 | 17.16 | 56.72 | 22.22 | 17.32 | 57 |
| 48 | M124Q | 3 | 0.21 | 1.33 | 2.00 | 21.25 | 17.01 | 55.04 | 22.22 | 18.03 | 58 |
| 49 | G129S | 1 | 0.11 | 1.34 | 2.00 | - | 0.00 | 0.00 | 22.21 | 8.41 | 53 |
| 50 | G129S | 2 | 0.11 | 1.36 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 51 | G129S | 3 | 0.11 | 1.37 | 2.00 | 21.85 | 0.07 | 0.45 | - | 0.00 | 0 |
| 52 | G129S | 1 | 0.21 | 1.39 | 2.00 | 21.74 | 0.62 | 2.01 | - | 0.00 | 0 |
| 53 | G129S | 2 | 0.21 | 1.37 | 2.00 | - | 0.00 | 0.00 | 22.48 | 0.06 | 0 |
| 54 | G129S | 3 | 0.21 | 1.39 | 2.00 | - | 0.00 | 0.00 | - | 0.00 | 0 |
| 55 | V88D/V97K | 1 | 0.10 | 1.37 | 0.50 | 21.07 | 13.13 | 359.54 | 21.96 | 1.95 | 53 |
| 56 | V88D/V97K | 2 | 0.10 | 1.17 | 0.50 | 21.06 | 19.45 | 454.67 | 21.88 | 0.69 | 16 |
| 57 | V88D/V97K | 3 | 0.10 | 1.42 | 0.50 | 21.13 | 9.38 | 266.25 | 21.89 | 1.89 | 54 |
| 58 | V88D/V97K | 1 | 0.20 | 1.42 | 0.50 | 21.03 | 19.24 | 273.07 | 21.88 | 0.51 | 7 |
| 59 | V88D/V97K | 2 | 0.20 | 1.42 | 0.50 | 21.08 | 21.28 | 302.02 | 21.91 | 0.75 | 11 |
| 60 | V88D/V97K | 3 | 0.20 | 1.42 | 0.50 | 21.11 | 20.66 | 293.22 | 21.93 | 0.52 | 7 |

**Table III.2 Average results for the *in vitro* folding analysis of *h* CD5d1 mutants**

| Construct | Protein conc. [mg/ml] | First peak RP-HPLC (native) | | | Second peak RP-HPLC(non-native) | |
|---|---|---|---|---|---|---|
| | | Average yield | Fraction of total soluble protein* (%) | Construct:WT | Average relative yield | Fraction of total soluble protein* (%) |
| Wild type | 0.1 | 62.4 | 49 | 1.00 | 65.9 | 51 |
| Wild type | 0.2 | 35.2 | 44 | 0.56 | 45.7 | 56 |
| L26K | 0.1 | 87.5 | 93 | 1.40 | 6.6 | 7 |
| L26K | 0.2 | 55.8 | 87 | 0.89 | 8.3 | 13 |
| W28R | 0.1 | 10.1 | 55 | 0.16 | 0.4 | 45 |
| W28R | 0.2 | 9.6 | 100 | 0.15 | 0.0 | 0 |
| V88D | 0.1 | 110.3 | 96 | 1.77 | 4.7 | 4 |
| V88D | 0.2 | 76.5 | 77 | 1.22 | 22.3 | 23 |
| V88N | 0.1 | 21.4 | 99 | 0.34 | 1.8 | 1 |
| V88N | 0.2 | 26.6 | 66 | 0.43 | 14.0 | 34 |
| V97K | 0.1 | 203.8 | 92 | 3.27 | 18.1 | 8 |
| V97K | 0.2 | 143.5 | 89 | 2.30 | 18.55 | 11 |
| M124N | 0.1 | 0.0 | - | 0.00 | 0.0 | 0 |
| M124N | 0.2 | 0.0 | - | 0.00 | 0.0 | 0 |
| M124Q | 0.1 | 87.0 | 58 | 1.39 | 62.6 | 42 |
| M124Q | 0.2 | 55.4 | 49 | 0.89 | 58.5 | 51 |
| G129S | 0.1 | 0.2 | 1 | 0.00 | 17.6 | 99 |
| G129S | 0.2 | 0.7 | 91 | 0.01 | 0.1 | 9 |
| V88D/V97K | 0.1 | 360.2 | 90 | 5.77 | 41.1 | 10 |
| V88D/V97K | 0.2 | 289.4 | 97 | 4.64 | 8.4 | 3 |

**Table III.3 Analysis of variance (ANOVA) for the *in vitro* folding analysis of *h* CD5d1 mutants (native peak)**

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | F value | p value | Interpretation |
|---|---|---|---|---|---|---|
| Protein concentration | 1 | 9379 | 9379 | 15.32 | $3.44 \times 10^{-4}$ | Significant |
| Construct | 9 | 544579 | 60509 | 98.86 | $1.81 \times 10^{-24}$ | Significant |
| Interaction | 9 | 9455 | 1051 | 1.72 | 0.12 | Not significant |
| Residual | 40 | 24482 | 612 | | | |
| Total | 59 | 587896 | | | | |
| Fisher's least significant difference (LSD) = 34.02 at 90% confidence interval | | | | | | |



**Table III.4 Analysis of variance (ANOVA) for the *in vitro* folding analysis of *h* CD5d1 mutants (non-native peak)**

| Source of variation | Degrees of freedom | Sum of squares | Mean squares | F value | p value | Interpretation |
|---|---|---|---|---|---|---|
| Protein concentration | 1 | 285 | 285 | 1.56 | 0.22 | Not significant |
| Construct | 9 | 256389 | 28488 | 155.99 | $2.87 \times 10^{-28}$ | Significant |
| Interaction | 9 | 769167.97 | 85463 | 467.97 | $1.27 \times 10^{-37}$ | Significant |
| Residual | 40 | 7305 | 183 | | | |
| Total | 59 | 1033147 | | | | |
| Fisher's least significant difference (LSD) = 18.60 at 90% confidence interval | | | | | | |

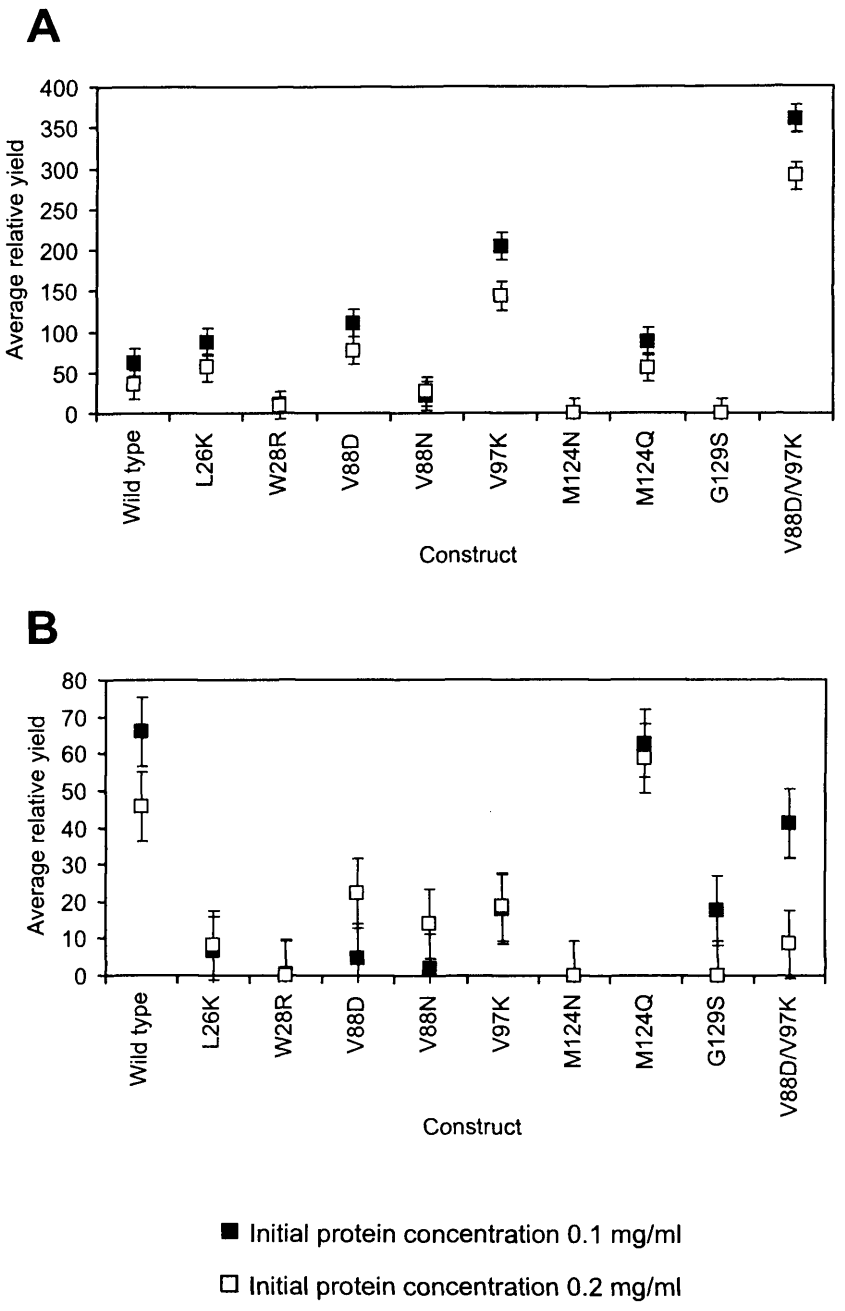


**Figure III.5** Average yields of native and non-native protein after *in vitro* folding of wild type and mutants of *h*CD5d1. (A) Native peak (B) Non-native peak.
The error bars extend LSD/2 above and below the average to display uncertainty intervals. Least significant difference (LSD) was calculated using the Fisher's method, *i.e.*
$LSD = t_{0.05}(v) * [2/3 * MSE]^{1/2}$
Where $t_{0.05}(v)$ is the critical value for the t-distribution at an upper-tail probability of 0.05 with v degrees of freedom. v denotes the degrees of freedom for the error in the ANOVA table (see Tables III.3 and III.4).
MSE is the mean square value for the error in the ANOVA table.

Mason RL, *et al.* (1989) *Statistical Design and Analysis of Experiments*, John Wiley & Sons.

*Kinetic observations on V88D/V97K hCD5d1*

The high yield obtained for the double mutant allowed further characterisation of the folding protocol. The *in vitro* folding of the double mutant *h*CD5d1 was monitored in order to investigate the time dependence of the yield of folded protein (Figure III.6). During the initial phase of folding the RP-HPLC trace progressively altered until ~25 hrs after which it was essentially invariant. At this point and up to 36 hours there was still a broad shoulder in the native peak. After dialysis, the shoulder diminished and became more highly resolved and it does not appear at all in the RP-HPLC chromatogram of the monomeric species after purification by SEC. Thus, the non-native shoulder was apparently formed by disulphide-bonded soluble oligomers such as those seen in lanes 1-3 of the non reducing SDS-PAGE of Figure III.3F. Non-reducing SDS-PAGE of the SEC fractions shown dimers, trimers and tetramers (Figure III.3F) that could be reduced with 2-mercaptoethanol (Figure III.3E). The resolution of the RP-HPLC chromatogram was not sufficient to separate the intermediate species during the folding reaction. As the non-native signal was very broad, many species must have been present during the folding process. One intermediate peak was sharper at 15 hours and it could represent a substantial population of a single stable intermediate.

*NMR spectroscopy*

V88D/V97K *h*CD5d1 could be concentrated to around of 1 mM without any significant precipitation in the same buffer previously used for WT *h*CD5d1 (50 mM potassium phosphate buffer pH 6.0, 0.2 M NaCl, 1 mM EDTA, 0.1% $NaN_3$). Thus, the double mutant has both remarkably better solubility and higher folding yield than the WT. The $^{1}H$-$^{15}N$ HSQC spectrum of V88D/V97K *h*CD5d1 has shown that the sample had a high concentration of a globular, folded component. Moreover, the distribution of signals was very similar that observed for the WT *h*CD5d1, suggesting that the mutations do not alter the overall fold of the domain (Figure III.7).

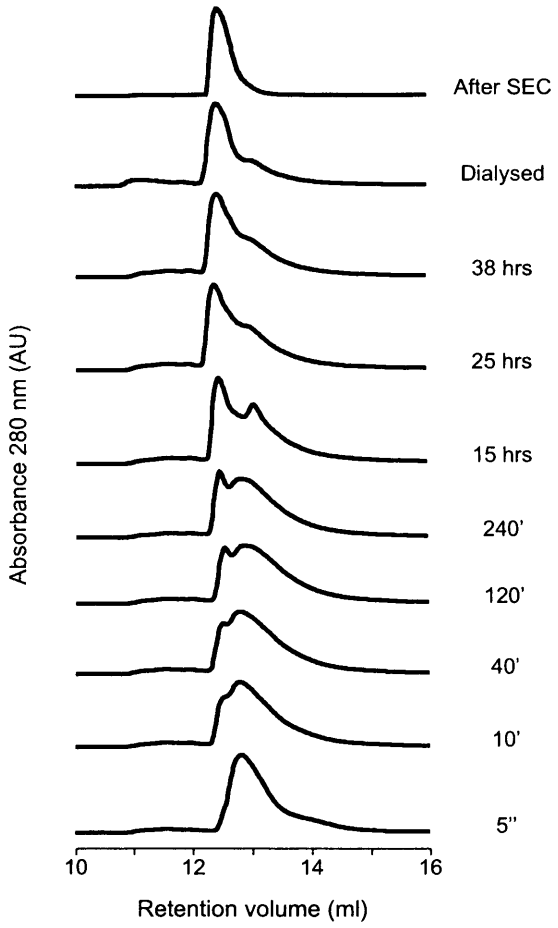


**Figure III.6** RP-HPLC elution profiles of the *in vitro* folding of V88D/V97K *h*CD5d1 at different time points. There is a progressive accumulation of the most hydrophilic species until 25 hours. This species is coincident with the protein monomer isolated by size-exclusion chromatography.

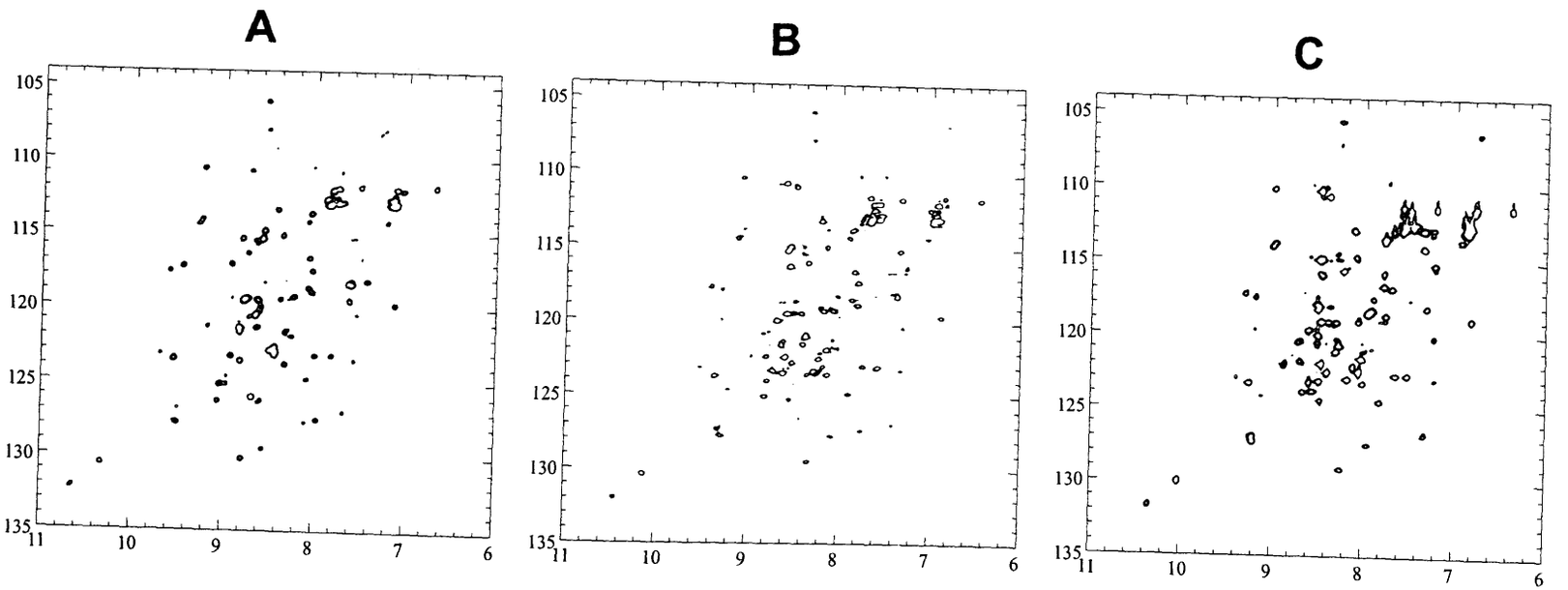


**Figure III.7** $^{1}H$- $^{15}N$ HSQC spectra at 298 K of different constructs of *h*CD5d1.
(A) Glycosylated *h*CD5d1 expressed in the yeast *Pichia pastoris* (Courtesy of Dr Mark Pfuhl).
(B) Unglycosylated *h*CD5d1expressed in *Escherichia coli.*
(C) Unglycosylated V88D/V97K *h*CD5d1 expressed in *Escherichia coli.*

## Discussion

The N-terminal domain of human CD5 expresses in eukaryotic systems with an N-linked glycan in residue N116[20]. The distribution of the signals in the $^{1}H$-$^{15}N$ HSQC spectrum of the glycosylated domain expressed in *Pichia pastoris*[21] is very similar to that obtained for the unglycosylated version presented in this work (Figure III.7), indicating that the presence of the N-linked carbohydrate is not essential for the attainment of the characteristic fold. However, the limited solubility of WT *h*CD5d1 expressed in bacteria could be a consequence of the lack of glycosylation.

It is generally accepted that one of the effects of the covalently bonded carbohydrates in proteins is to increase protein stability or solubility, or both, and to prevent aggregation. A statistical study of the glycosylated proteins in the PDB has shown that the regions of the protein surface that are in contact with glycans are more hydrophobic than the rest of the protein surface[22]. This is in agreement with experimental work suggesting that complex carbohydrates improve solubility because they shield solvent-exposed hydrophobic patches[23-26]. Other experimental evidence indicates that the glycans increase protein solubility because of their inherent polar properties, as there appears to be linear correlation between solubility and number of carbohydrate residues[27].

Glycans tend to be chemically heterogeneous and seldom rigid, and as a consequence glycoproteins are generally difficult to crystallise[28]. The carbohydrate also increases the molecular weight, so glycoproteins are not often suitable for NMR spectroscopy. Because the three-dimensional structure of a protein is not generally influenced by the glycan[29], it is common to circumvent these problems by removing most of the carbohydrate by chemical or enzymatic digestion[21,30-32], or by mutating the glycosylation site to Gln[33,34].

In general, bacterial expression systems are unable to produce glycosylated proteins. Nevertheless, because of their many merits, and the complications of working with glycosylated proteins, it is worth attempting the expression of a glycan-free form of structural targets. Unglycosylated proteins expressed in *E. coli* are likely to be active[8,10,35], however, they are often less soluble and prone to aggregation, a problem that has been solved in some instances by introducing mutations that decrease the hydrophobicity of the protein surface[8,10,36]. In this Chapter it was shown that both the solubility and the *in vitro* folding yield of the unglycosylated *h*CD5d1 could be enhanced by replacing non-conserved hydrophobic residues.

The influence of the different mutations is difficult to interpret as they could have altered, at least, three different phenomena: (1) the solubility of the extended polypeptide chain; (2) the solubility of the folded conformation; and (3) the stability of intermediate folding states. There is also the possibility that some of the detrimental mutations prevented the protein from folding because they disturb the side chain packing. In order to assess whether there was any indication of relationship between the properties of the sequence and the folding yield, the Kyte-Doolittle hydrophobicity[37], the Hopp-Woods hydrophilicity[38], and the Rost and Sander predicted solvent accessibility[39] were computed for the different positions of the WT where mutations were introduced (Figure III.8A,B,C). No connection was found between the success of a particular mutation and the predicted solvent accessibility or the Kyte-Doolittle hydrophobicity (Spearman correlation coefficients[2], $\rho$, of -0.09 and -0.03 respectively). The correlation score for the Hopp-Woods values was much higher ($\rho$ = -0.59), this value indicating a significant correlation with a confidence level of 80%. There was also no correlation between the calculated side-chain solvent accessibility[15] for the equivalent residues in the M2BP SRCR structure (Figure III.8C). However, this may simply mean that the structural context of these residues is different in the *h*CD5d1 and M2BP SRCR domains. Average values of hydropathy index (GRAVY) using the Kyte-Doolittle scores were calculated with ExPASy's[40] Protparam for each mutant. The Spearman value in this case was high ($\rho$ =-0.72), pointing out a relationship between the GRAVY score and the folding yield

[2] Spearman correlation coefficient ($\rho$) is a better indicator of the existence of a relationship between two variables when this relationship is not linear. Like the least squares correlation coefficient, a $\rho$ value of $\pm 1$ indicates a perfect relationship, and a value of 0 indicates no relationship.

with a significance level of 95%. GRAVY indices depend only on the amino acid sequence, not the 3D structure. If this correlation is real, it could mean that the unfolded WT *h*CD5d1 falls out of solution because of the intrinsic hydrophobicity of the amino acid sequence, while the *h*CD5d1 mutants with very low or no folding yields precipitate in later intermediate states or are unable to fold because the three-dimensional structure has been disturbed.

Clearly, the strategy of mutation of potentially solvent-exposed residues has been successful in modifying the solubility of *h*CD5d1. However, the location of the mutations was not the only determinant for the dissimilar properties found for the mutants. For a given site substitution of different amino acids resulted in different behaviours. In the case of V88, replacement for Asp was very successful, but not for Asn. Perhaps the difference was just a consequence of Asp being more hydrophilic than Asn. In the case of M124 the substitution for Asn prevented the protein from folding, while replacement for Gln resulted in a higher *in vitro* folding yield than WT. At a purely speculative level, this result could be a consequence of the difference in side chain size. Perhaps the methylene group in the γ position, present in Gln and Met but not in Asn, is important for adequate side chain packing.

For *h*CD5d1, higher folding yields accompanied lower levels of precipitation during *in vitro* folding. Apparently the successful mutations were those that prevented aggregation of the unfolded species and/or folding intermediates. While V88D/V97K *h*CD5d1 has improved solubility and *in vitro* folding yields, these two properties may not always be related. Studies with disulphide bonded proteins expressed in the bacterial periplasm have shown that *in vivo* folding yields correlate with aggregation rates but not with the solubility of the purified protein[41]. Other studies have found a direct correlation between high thermodynamic stability of the native structure and high *in vivo* folding yields[42].

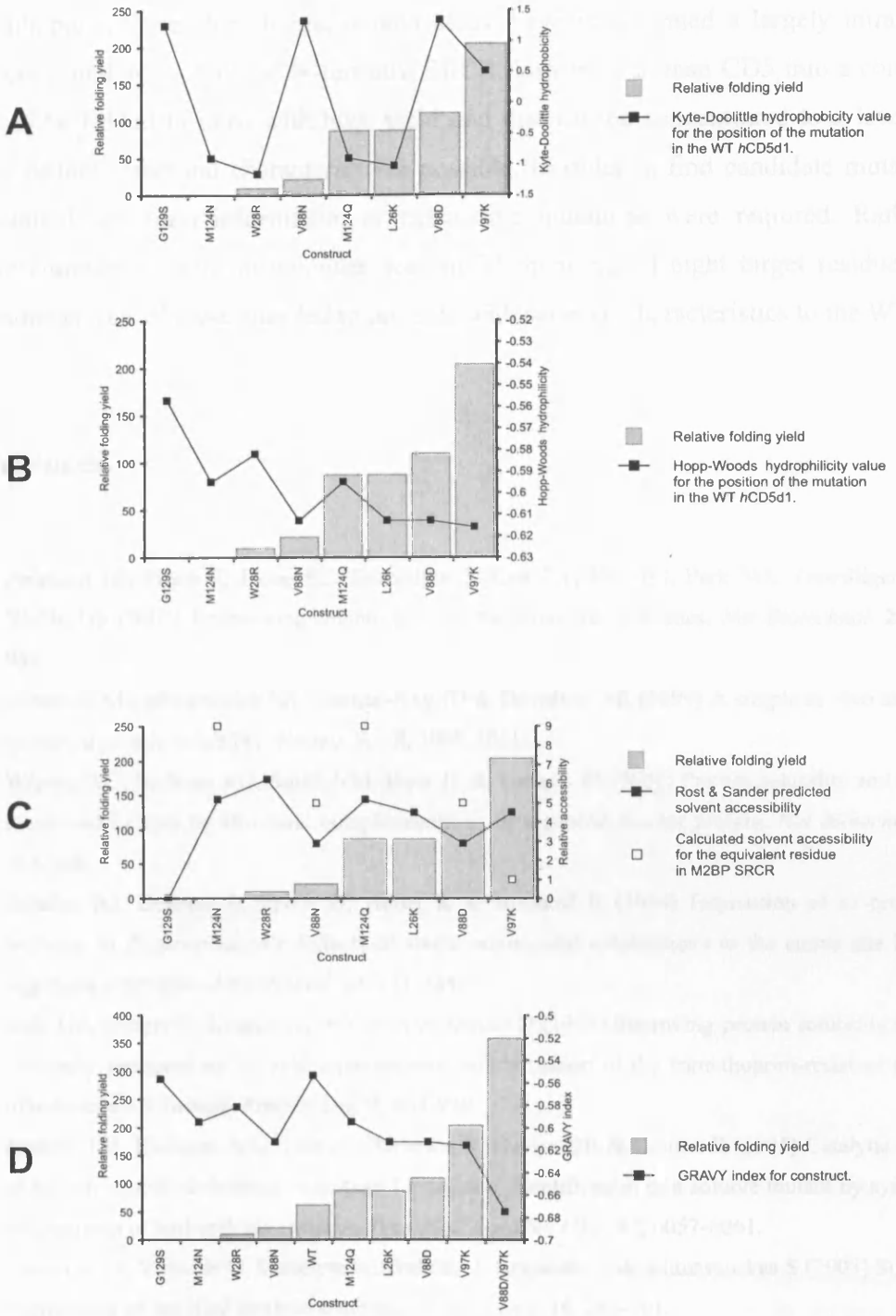


**Figure III.8** Comparison of relative refolding yields of *h*CD5d1 variants with different physical parameters. (A) Kyle-Doolittle hydrophobicity score for the position of the mutation in the wild type construct using a window size of 9. The higher the score, the higher the hydrophobicity of the position.(B) Hopp-Woods hydrophilicity score for the position of the mutation in the wild type construct using a window size of 7. The higher the score the higher the hydrophilicity of the position. (C) Predicted solvent accessibility using the Rost and Sander algorithm and calculated solvent accessibility of the corresponding residue side chain in M2BP SRCR structure, if available. (D) Grand average of hydropathicity (GRAVY) for each of the mutants.

To summarise, two replacements of non-conserved, and presumably solvent exposed, hydrophobic residues for charged amino acids, have transformed a largely intractable unglycosylated version of the N-terminal SRCR domain of human CD5 into a construct that can be folded *in vitro* with high yield and that can be concentrated to a level that makes further structural characterisation possible. In order to find candidate mutations, no detailed structural information or exhaustive mutations were required. Rather, a simple comparison with orthologues was sufficient to select eight target residues and mutations at four of these sites led to proteins with superior characteristics to the WT.

## References


1. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC & Waldo GS (2002) Engineering soluble proteins for structural genomics. *Nat Biotechnol.* **20**, 927-932.
2. fffaxwell KL, Mittermaier AK, Forman-Kay JD & Davidson AR (1999) A simple in vivo assay for increased protein solubility. *Protein Sci.* **8**, 1908-1911.
3. Wigley WC, Stidham RD, Smith NM, Hunt JF & Thomas PJ (2001) Protein solubility and folding monitored in vivo by structural complementation of a genetic marker protein. *Nat Biotechnol.* **19**, 131-136.
4. Schulze AJ, Degryse E, Speck D, Huber R & Bischoff R (1994) Expression of à1-proteinase inhibitor in *Escherichia coli*: Effects of single amino acid substitutions in the active site loop on aggregate formation. *J Biotechnol.* **32**, 231-238.
5. Dale GE, Broger C, Langen H, D'Arcy A & Stuber D (1994) Improving protein solubility through rationally designed amino acid replacements: solubilization of the trimethoprim-resistant type S1 dihydrofolate reductase. *Protein Eng.* **7**, 933-939.
6. Jenkins TM, Hickman AB, Dyda F, Ghirlando R, Davies DR & Craigie R (1995) Catalytic domain of human immunodeficiency virus type 1 integrase: identification of a soluble mutant by systematic replacement of hydrophobic residues. *Proc Natl Acad Sci USA.* **92**, 6057-6061.
7. Daujotyte D, Vilkaitis G, Manelyte L, Skalicky J, Szyperski T & Klimasauskas S (2003) Solubility engineering of the HhaI methyltransferase. *Protein Eng.* **16**, 295-301.
8. Sun ZY, Dotsch V, Kim M, Li J, Reinherz EL & Wagner G (1999) Functional glycan-free adhesion domain of human cell surface receptor CD58: design, production and NMR studies. *EMBO J.* **18**, 2941-2949.
9. McElroy HE, Sisson GW, Schoettlin WE, Aust RM & Villafranca JE (1992) Studies on engineering crystallizability by mutation of surface residues of human thymidylate synthase. *J Cryst Growth.* **122**, 265-272.

10. Malissard M & Berger EG (2001) Improving solubility of catalytic domain of human beta-1,4-galactosyltransferase 1 through rationally designed amino acid replacements. *Eur J Biochem.* **268**, 4352-4358.
11. Das D & Georgiadis MM (2001) A directed approach to improving the solubility of Moloney murine leukemia virus reverse transcriptase. *Protein Sci.* **10**, 1936-1941.
12. Mosavi LK & Peng ZY (2003) Structure-based substitutions for increased solubility of a designed protein. *Protein Eng.* **16**, 739-745.
13. Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
14. Hohenester E, Sasaki T & Timpl R (1999) Crystal structure of a scavenger receptor cysteine-rich domain sheds light on an ancient superfamily. *Nat Struct Biol.* **6**, 228-232.
15. Hubbard, S. J. and Thornton, J. M. (1993) NACCESS (version 2.1.1) University College London.
16. Gill SC & von Hippel PH (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem.* **182**, 319-326.
17. Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL & Sykes BD (1995) $^{1}H$, $^{13}C$ and $^{15}N$ chemical shift referencing in biomolecular NMR. *J Biomol NMR.* **6**, 135-140.
18. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J & Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* **6**, 277-293.
19. Boucher, W (2002) AZARA suite for NMR data (version 2.7) Department of Biochemistry, University of Cambridge.
20. McAlister MS, Brown MH, Willis AC, Rudd PM, Harvey DJ, Aplin R, Shotton DM, Dwek RA, Barclay AN & Driscoll PC (1998) Structural analysis of the CD5 antigen -expression, disulphide bond analysis and physical characterisation of CD5 scavenger receptor superfamily domain 1. *Eur J Biochem.* **257**, 131-141.
21. McAlister MS, Davis B, Pfuhl M & Driscoll PC (1998) NMR analysis of the N-terminal SRCR domain of human CD5: engineering of a glycoprotein for superior characteristics in NMR experiments. *Protein Eng.* **11**, 847-853.
22. Petrescu AJ, Milac AL, Petrescu SM, Dwek RA & Wormald MR (2004) Statistical analysis of the protein environment of N- glycosylation sites: implications for occupancy, structure, and folding. *Glycobiology.* **14**, 103-114.
23. Toyoda T, Arakawa T & Yamaguchi H (2002) N-glycans stabilize human erythropoietin through hydrophobic interactions with the hydrophobic protein surface: studies by surface plasmon resonance analysis. *J Biochem (Tokyo).* **131**, 511-515.
24. Song Y, Azakami H, Hamasu M & Kato A (2001) *In vivo* glycosylation suppresses the aggregation of amyloidogenic hen egg white lysozymes expressed in yeast. *FEBS Lett.* **491**, 63-66.
25. Erbel PJ, Karimi-Nejad Y, van Kuik JA, Boelens R, Kamerling JP & Vliegenthart JF (2000) Effects of the N-linked glycans on the 3D structure of the free alpha-subunit of human chorionic gonadotropin. *Biochem.* **39**, 6012-6021.
26. Ioannou YA, Zeidner KM, Grace ME & Desnick RJ (1998) Human α-galactosidase A: glycosylation site 3 is essential for enzyme solubility. *Biochem J.* **332**, 789-797.

27. Tams JW, Vind J & Welinder KG (1999) Adapting protein solubility by glycosylation. N-glycosylation mutants of *Coprinus cinereus* peroxidase in salt and organic solutions. *Biochim Biophys Acta.* **1432**, 214-221.
28. Wormald MR & Dwek RA (1999) Glycoproteins: glycan presentation and protein-fold stability. *Structure Fold Des.* **7**, R155-R160.
29. Helenius A & Aebi M (2001) Intracellular functions of N-linked glycans. *Science.* **291** , 2364-2369.
30. Rydberg EH, Sidhu G, Vo HC, Hewitt J, Cote HC, Wang Y, Numao S, MacGillivray RT, Overall CM, Brayer GD & Withers SG (1999) Cloning, mutagenesis, and structural analysis of human pancreatic α-amylase expressed in *Pichia pastoris*. *Protein Sci.* **8**, 635-643.
31. Gordon K, Redelinghuys P, Schwager SL, Ehlers MR, Papageorgiou AC, Natesh R, Acharya KR & Sturrock ED (2003) Deglycosylation, processing and crystallization of human testis angiotensin-converting enzyme. *Biochem J.* **371**, 437-442.
32. Avramopoulou V, Mamalaki A & Tzartos SJ (2004) Soluble, oligomeric and ligand-binding extracellular domain of human alpha7 acetylcholine receptor expressed in yeast. Replacement of the hydrophobic Cys-loop by the hydrophilic loop of ACh-binding protein enhances protein solubility. *J Biol Chem.* **279**, 38287-38293.
33. Bruinzeel W, Yon J, Giovannelli S & Masure S (2002) Recombinant insect cell expression and purification of human á-secretase (BACE-1) for X-ray crystallography. *Protein Expr Purif.* **26**, 139-148.
34. Bowman KK, Clark J, Yu L, Mortara K, Radika K, Wang J & Zhan H (2000) Expression, purification, and characterization of deglycosylated human pro-prostate-specific antigen. *Protein Expr Purif.* **20**, 405-413.
35. Wenig K & Sondermann P (2003) Purification, crystallization and X-ray diffraction analysis of the extracellular part of the human Fc receptor for IgA, FcàRI (CD89). *Acta Crystallogr D Biol Crystallogr.* **59**, 2247-2250.
36. Narhi LO, Arakawa T, Aoki K, Wen J, Elliott S, Boone T & Cheetham J (2001) Asn to Lys mutations at three sites which are N-glycosylated in the mammalian protein decrease the aggregation of Escherichia coli-derived erythropoietin. *Protein Eng.* **14**, 135-140.
37. Kyte J & Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* **157**, 105-132.
38. Hopp TP & Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci USA.* **78**, 3824-3828.
39. Rost B & Sander C (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins.* **20**, 216-226.
40. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD & Bairoch A (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784-3788.
41. Nieba L, Honegger A, Krebber C & Pluckthun A (1997) Disrupting the hydrophobic patches at the antibody variable/constant domain interface: improved in vivo folding and physical characterization of an engineered scFv fragment. *Protein Eng.* **10**, 435-444.
42. Chan W, Helms LR, Brooks I, Lee G, Ngola S, McNulty D, Maleeff B, Hensley P & Wetzel R (1996) Mutational effects on inclusion body formation in the periplasmic expression of the immunoglobulin VL domain REI. *Fold Des.* **1**, 77-89.

# Chapter Four

# The NMR solution structure of the N-terminal SRCR domain of human CD5


## Abstract

The results of nuclear magnetic resonance (NMR) experiments performed on the N-terminal domain of human CD5 at 298 K revealed that regions of the structure are mobile on the microsecond to millisecond (intermediate) timescale, which is manifest as non-uniform peak broadening and irregular signal intensities in the $^{1}H$-$^{15}N$ heteronuclear single quantum coherence (HSQC) spectrum. Increasing the temperature above 298 K was found to greatly enhance the quality of the $^{1}H$-$^{15}N$ HSQC spectrum. Three-dimensional NMR experiments at 318 K enabled the assignment of more than 95% of the main-chain and side chain resonances. The solution structure of the N-terminal domain of human CD5 was determined at 318 K using standard interproton distance and dihedral angle-restrained molecular dynamics protocols. Forty percent of the residues were found to be in structurally well-defined regions, including all of the regular secondary structure features found for other members of the scavenger receptor cysteine-rich (SRCR) superfamily. The remaining residues of the polypeptide appear to be distinctly less well ordered. Comparison with previous data obtained for a glycosylated form of the CD5d1 domain suggests that some of the mobile residues are flexible because of the absence of the carbohydrate. However, some residues are mobile in both the glycosylated and the unglycosylated domains, suggesting that there is an inherent mobility in specific regions of the domain that may have functional significance.

Chapter abbreviations and acronyms

***h*CD5d1** N-terminal SRCR domain of CD5

***h*CD5d1m** V88D/V97K *h*CD5d1

**CD6d3** C-terminal SRCR domain of leukocyte receptor CD6

**HSQC** heteronuclear single quantum coherence

**M2BP** Mac-2 binding protein

**NMR** nuclear magnetic resonance

**NOE** nuclear Overhauser effect

**PDB** Protein Data Bank

**ppm** parts per million

**RMSD** root mean square deviation

**SRCR** scavenger receptor cysteine-rich

## Introduction

The scavenger receptor cysteine-rich (SRCR) domain is a globular disulphide-bonded α and β fold of around 110 residues. SRCR domains are almost exclusively found in secreted and/or membrane-anchored metazoan proteins and are usually glycosylated. A few SRCR domains have also been found in some members of the Apicomplexa parasite family (*Plasmodium* and *Toxoplasma* genera)[1,2], probably indicating a horizontal gene transfer event. Proteins with SRCR domains are expressed in a great variety of cell types, but predominantly in cells of the immune system, epithelial cells, hepatocytes and neurones.

The SRCR domain superfamily was discovered during the analysis of the amino acid sequence of the mammalian type-II macrophage scavenger receptor type A (SR-A)[3]. SR-As are transmembrane glycoproteins that bind a broad range of polyanions, including modified low density lipoproteins, phospholipids and polynucleotides. However, the SRCR domain of the SR-A does not appear to be involved in the polyanion recognition, as the absence of this domain does not compromise ligand binding[4].

The number of cysteine residues in an SRCR domain varies from four to eight. When eight cysteines are present, the disulphide bond connectivity is C1-C4, C2-C7, C3-C8 and C5-C6[5] (where the numbering refers to the linear order of occurrence of the cysteine

residues in the amino acid sequence of the protein), as has been confirmed experimentally for the N-terminal SRCR domain of CD5[6] and corroborated by the two solved structures from the superfamily[7,8]. Every known SRCR domain has at least two putative intradomain disulphide bonds, one of which is always equivalent to C3-C8. Traditionally, the superfamily had been divided in two groups, A and B, according to the number of cysteines present[9]. However, this criterion has now been found to be inadequate as new sequences with different numbers of cysteines have been incorporated into the family. A more robust classification divides group A and B according to their genomic organisation: Group A domains are encoded by two exons and group B members are encoded in a single exon[10]. The members of each group have varying numbers of cysteines, but no group A member has been reported with a predicted C1-C4 disulphide bond.

The proteins of the superfamily are either formed by several tandem copies of the SRCR domain or are mosaic proteins (Figures IV.1-3). So far, 25 different types of domain have been found together with SRCR domains. Of these associated domains, 22 are protein-, lipid- or carbohydrate- binding, and three have enzymatic activity (Table IV.4). SRCR domains of group B have only been found in association with six of these domains (in the same protein), all of them implicated in protein- and/or sugar- binding. Group A SRCR domains are less exclusive and occur in association with 14 different domain types. In addition to protein-, lipid- and/or carbohydrate- binding domains, group A SRCR domains have also been found in association with enzymatic domains, namely tyrosine kinases, serine proteases and lysyl oxidases.

Some SRCR domains themselves are involved in protein-protein interactions. The only SRCR binding event that has been studied in any detail is the interaction of the membrane-proximal SRCR domain of CD6 (CD6d3) with one of the immunoglobulin C-type domains of CD166 (also known as the activated leukocyte-cell adhesion molecule, ALCAM). Other protein-protein interactions that probably involve SRCR domains, but have not been studied in detail, are CD163 binding to the haptoglobin-haemoglobin complex[11], gp340 binding to the C-lectin domains of surfactant proteins A and D[12,13], and the interaction of gp340 with the sphingolipid binding motif (V3 loop) of gp120[14] (a human immunodeficiency virus-1 surface envelope protein). Many other interactions of SRCR superfamily proteins have been mapped to regions outside the SRCR domain.

**Table IV.1 Non-Chordata members of the scavenger receptor cysteine-rich superfamily group A**

| Name | | Uniprot entry | Number of SRCR domains | Other domains‡ | Species and/or tissue distribution | Function | Transmembrane or secreted | Reference |
|---|---|---|---|---|---|---|---|---|
| SRCR Tyrosine Kinase* | | O97191 | 1 + 1 truncated | TK (1) | | Unknown | T | TrEMBL |
| Putative gene | Sponge aggregation receptor SRCR-SCR-Car | O96943 | 14 | CCP (6) | Marine sponge (*G. cydonium*) | Cell aggregation receptor | T | (15) |
| | SRCR-Re | | 12 | CCP (1) | | Cell-cell interaction | T | (16) |
| | SRCR-Mo | | 10 | None | | | S | |
| SRCR_SD | | Q8SSX3 | 4 | None | Marine sponge (*S.domuncula*) | Cell aggregation | T | (17) |
| SP85 / Receptor for sperm-activating peptide (Speract/SAP-I) | | P16264 Q17064 Q25111 | 4 | None | Sea urchin spermatozoa flagella (*A. punctulata, S.purpuratus H. pulcherrimus*) | Binds decapeptide secreted by eggs for spermatozoa motility | T | (18)(19) |
| Scavenger receptor cysteine rich protein variant 1 (SpSRCR1) | | O97378 | 4 | TIL (4) | Sea urchin coelomocytes (*S. purpuratus*) | Innate immune response | S | (20) |
| Scavenger receptor cysteine-rich protein variant 5 (SpSRCR5) | | O97379 | 2 | CUB (2) | | | S | (20) |
| Scavenger receptor cysteine-rich protein type 6 (SpSRCR6) | | Q9NC91 | 3 | TIL (1) | | | S | (21) |
| Scavenger receptor cysteine-rich protein type 7.1 (SpSRCR7.1) | | Q9NC90 | 7 | EGF (3), CCP (1) | | | T | (21) |
| Scavenger receptor cysteine-rich protein type 7.2 (SpSRCR7.2) | | Q9NC89 | 7 | CCP (1) | | | T | (21) |
| Scavenger receptor cysteine-rich protein type 12 (SpSRCR12) | | O97375 | 20 | None | | | S | (21) |
| Scavenger receptor cysteine-rich protein type 20 (SpSRCR20) | | Q9NC88 | 7 | None | | | S | (21) |
| c06b8.7* | | O17575 Q9VQV1 Q7PVZ0 | 2/3 | PBH (20/21), CTL (1) | *C. elegans*, Insecta | - | T | PFAM SMART |
| Tequila gene products | SP22D / GRAAL* / Adhesive serine protease | Q9U1I2 Q9NJS5 | 2 | CBD2 (2), LDL (2), SP (1) | Insect immune-related cell types, hemolymph | Innate immune response | T/S | (22) |
| | GRAAL2* | Q9BI19 | 2 | CBD2 (15), LDL (2), SP (1) | *D. melanogaster* | Upregulated in immune challenge | S | PFAM SMART |
| cg2105-pa* (Corin) | | Q9V4N6 | 1 | FZ (1), LDL (2), SP (1) | | | S | (23) |
| Lysyl oxidase-like | | Q9V9X5 | 1 | LOXL (1) | | | - | (24) |
| Lysyl oxidase-like 2 | | Q9N9Y8 Q7QBA0 | 2 | LOXL (1) | Insecta | - | - | PFAM, SMART |
| PSLAP/PxSR | | Q8IM24 | 2 | LH2 (1) LCCL(4) | *Plasmodium* spp. *Toxoplasma* spp. | - | S | (1)(2) |

*Hypothetical protein; ‡ Please refer to Table IV.4 for information about these domains

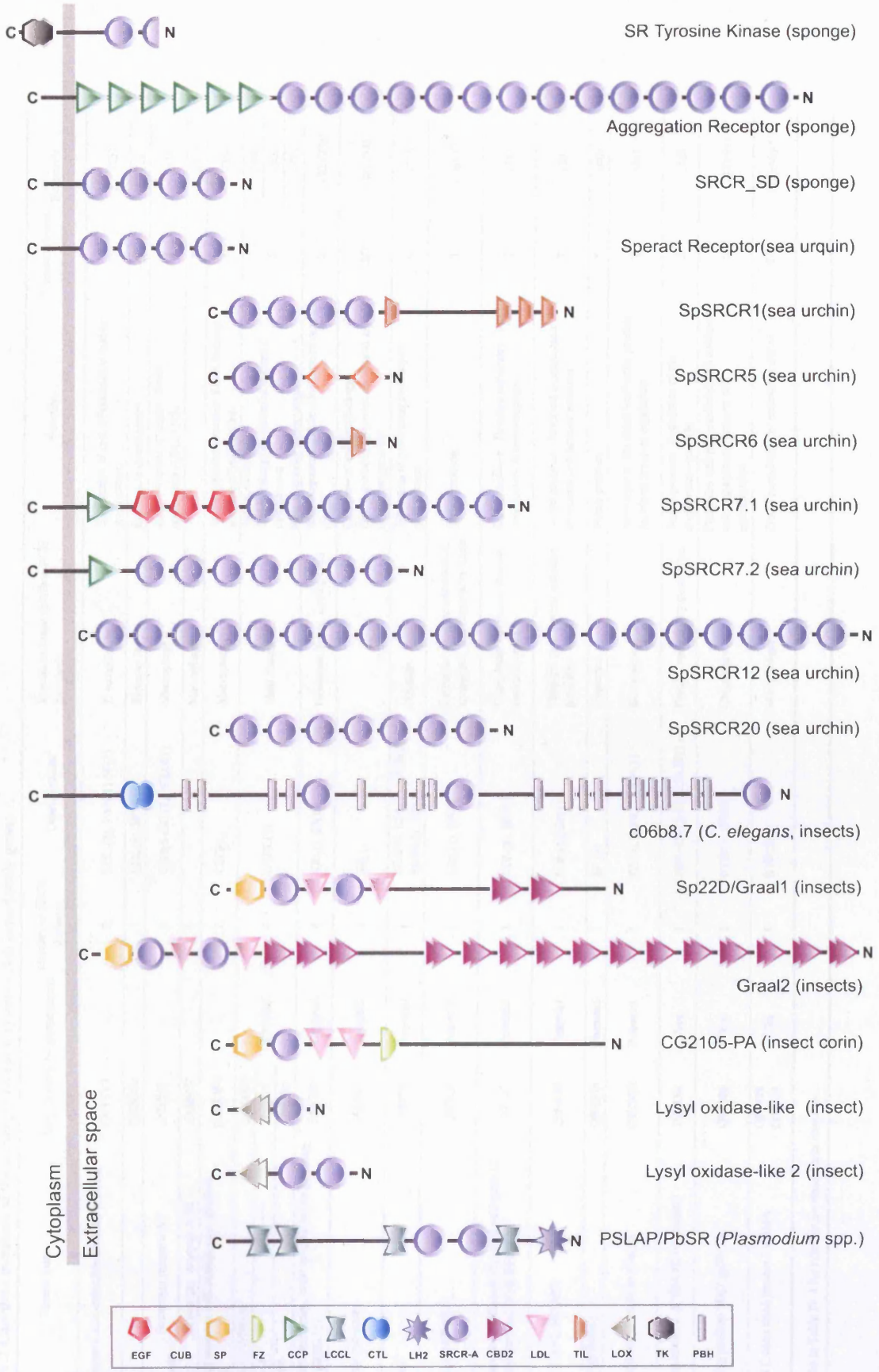


**Figure IV.1** Domain organisation of the non-Chordata protein members of the scavenger receptor cysteine-rich (SRCR) superfamily (refer to Tables IV.1 and IV.4 for details and abbreviations).

**Table IV.2 Chordata members of the scavenger receptor cysteine-rich superfamily group A**

| Protein name | Uniprot entry | N- glycosylation | Number of SRCR domains | Other domains‡ | Species or tissue distribution in mammals | Function | Transmembrane or secreted | Reference |
|---|---|---|---|---|---|---|---|---|
| Tunicate retinoic acid-inducible modular protease (TRAMP) | Q9Y1V3 | - | 2 | LDL (3), PAN(1) SP(1) | *T. misakiensis* | Stimulation of cell differentiation and/or proliferation | S | (25) |
| Xesp-2 | Q9DGR2 | - | 1 | LDL(8), SP (1) | *Xenopus laevis* | Embryonic development | T | (26) |
| SR-A gene: Scavenger receptor A I | P21757 | - | 1 | Coiled-coil (1), CLG(1) | Macrophage | Innate immune response. Binds polyanions (LDL, LPS). | T | (3) |
| SR-A gene: Scavenger receptor A III | O60505 | - | 1 | - | Macrophage | - | T | (27) |
| Macrophage receptor with collagenous structure (MARCO) | Q9UEW3 | - | 1 | CLG(1) | Macrophage | Innate immune response. Binds bacteria and lipopolysacharides | T | (28) |
| LOXL-like 2 / WS914 | Q9Y4K0 | Potential | 4 | LOXL(1) | Most tissues | Amine oxidase | S | (29) |
| LOXL-like 3 | P58215 | | | | | Cross-linking of extracellular matrix components | | (30) |
| LOXL-like 4 | Q96JB6 | | | | | Role in tumour suppression | | (31) |
| Neurotrypsin (motopsin, brain specific serine protease 3 , leydin, PRSS-12) | P56730 | Potential | 4 | KR(1), SP(1). | Neurones, lungs, Leydig cells | Morphogenesis and development of neural tissue | S | (32)(33) |
| Hepsin serine protease | P05981 | Potential | 1 | SP(1) | - | Cell growth and morphology. Overexpression promotes metastasis and cancer progression | S/T | (8) (34) |
| Enterokinase | P98073 | Potential | 1 | SEA(1), LDL(2), CUB(2), MAN(1) , SP(1) | Intestine | Activation of pancreatic proteolytic proenzymes | S | (35) |
| Epitheliasin (TMPRSS2) | Q9JIQ8 | Potential | 1 | LDL(1), SP(1) | Epithelia of the gastrointestinal, urogenital, and respiratory tracts | Serine protease | T | (36)(37) |
| Tumour associated differentially-expressed gene-12 protein serine protease (TMPRSS3) | P57727 | Potential | 1 | LDL(1), SP(1) | Wide. Inner ear tissue, thymus, stomach, testis | Serine protease. Possible substrates: Ion transporters & neurotrophins | T | (38) |
| TMPRSS4 (CAP2, MP-SP2) | Q9NRS4 | Potential | 1 | LDL(1), SP(1) | Traquea, colon, small intestine, pancreas | Serine protease. Involved in metastasis formation and tumour invasion | T | (39) |
| Spinesin (TMPRSS5) | Q9H3S3 | Potential | 1 | SP (1) | Neurones | Serine protease | T | (40) |
| Heart specific serine protease (Corin) | Q9Y5Q5 | Potential | 1 | FZ (2), LDL(7), SP (1) | Heart myocytes | Activation of the atrial natriuretic peptide for blood pressure regulation | T | (41) |
| Complement Factor I (C3B/C4B inactivator) | P05156 | Yes | 1 | FIMAC (1), LDL(2),SP(1) | Plasma, secreted by monocytes | Serine protease. Regulation of the complement cascade | S | (42) |
| Mac2 Binding protein/ 90K/ gp90/ L3 | Q08380 | Yes | 1 | BTB(1), IVR(1) | Ubiquitous | Promotes intergrin-mediated cell adhesion and aggregation. Tumour cell embolisation | S | (43)(44) |
| Cyclophilin C-associated protein / MAMA | O35649 O70513 | Yes | 1 | BTB(1) | Macrophages | Down modulation of endotoxin signal | T/S | (45)(46)(47) |

‡ Please refer to Table IV.4 for information about these domains

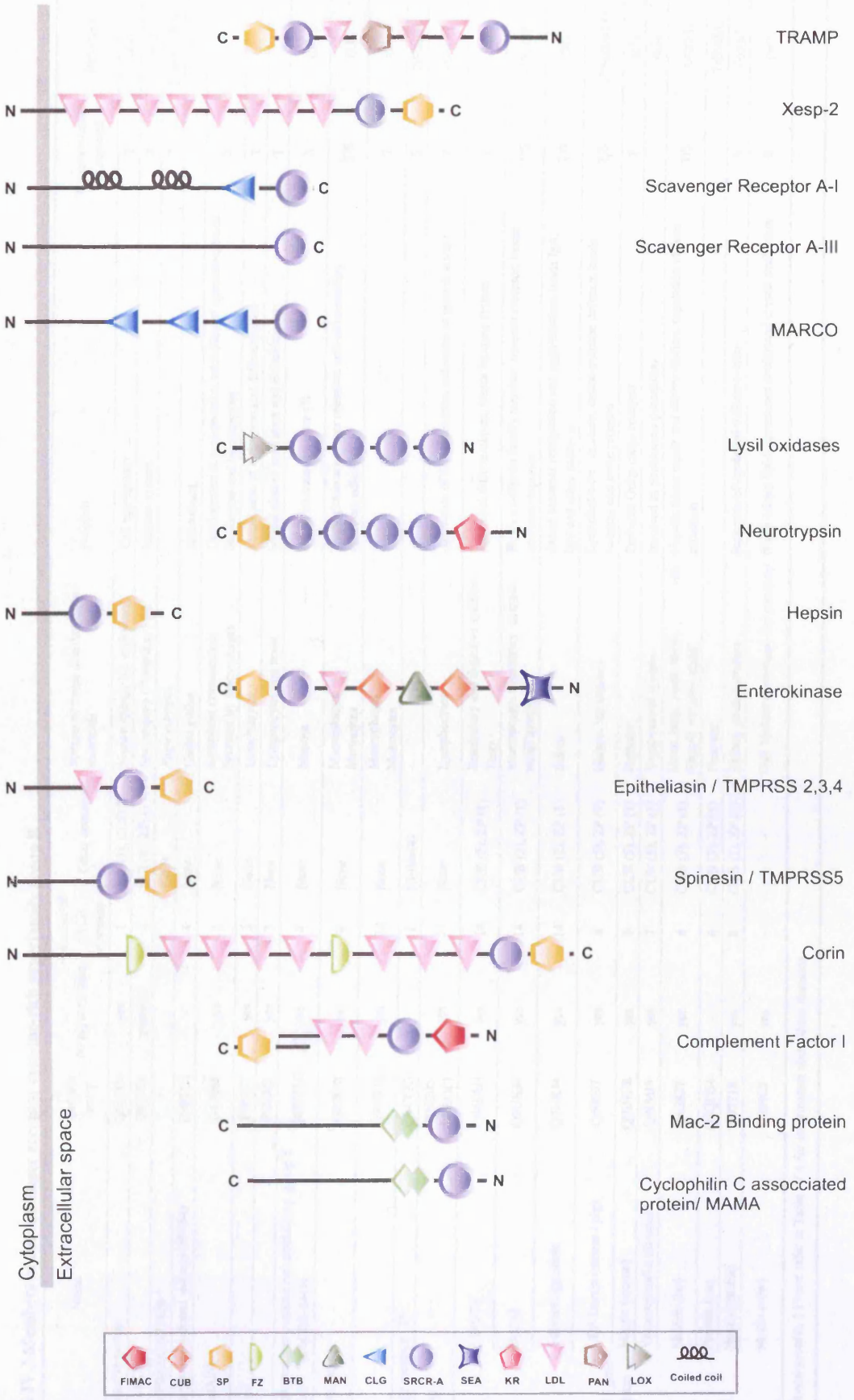


**Figure IV.2** Domain organisation of the Chordata protein members of the scavenger receptor cysteine-rich (SRCR) superfamily group A (refer to Tables IV.2 and IV.4 for details and abbreviations).

**Table IV.3 Members of the scavenger receptor cysteine-rich superfamily group B**

| Name | | Uniprot entry | N- glycosylation | Number of SRCR domains | Other domains‡ | Species or tissue distribution in mammals | Function | Transmembrane or secreted | Reference |
|---|---|---|---|---|---|---|---|---|---|
| Multiadhesive protein | | Q967D4 | yes | 1 | FN3(2), CCP(1) | Marine sponge (*G. cydonium*) | Cell aggregation | T | (48) |
| PEMA-SRCR | | Q92098 | potential | 2 | EGF(4), ZP(1) | Sea lamprey (*P.marinus*) gut | Immune system | T | (49) |
| SINFRUP000000137426 * | | - | - | | IG(2) | *Fugu rubripes* | - | T | Ensembl Fugu |
| 18-B Turpentine-induced protein (PIT54) | | Q98TD1 | - | 4 | None | *Gallus gallus* | Antioxidant | | (50) |
| Sp-alpha/AIM | | O43866 | yes | 3 | None | Lymphoid compartment<br>Secreted by macrophages | Development & maintenance; inhibition of apoptosis; binds monocytes and lymphocytes | S | (51) |
| CD5 | | P06127 | yes | 3 | None | Lymphocytes | Modulation of activation and differentiation | T | - |
| CD6 | | P30203 | yes | 3 | None | Lymphocytes and brain | Modulation of activation and differentiation | T | - |
| Scavenger receptor cysteine rich containing, group B 4 domains (SRCRB-S4D) | | Q8WTU2 | no | 4 | None | Mucosa | Innate immune response (?) | S | (52) |
| CD163/M130 | | Q07898 | yes | 9 | None | Macrophages<br>Monocytes | Hemoglobin scavenger receptor; anti-inflammatory function; adhesion | T/S | (11) |
| M160 | | Q9NR16 | yes | 12 | None | Macrophages<br>Monocytes | - | T | (53) |
| Scavenger receptor 2* | | Q91YK7 | | 1 | IG-like(4) | - | - | S | TrEMBL |
| WC1/T19 | | P30205<br>Q9BDF1 | yes | 11 | None | Lymphocytes | Activation of differentiation; induction of growth arrest | T | (54)(55) |
| DMBT1 gene products | DMBT1 | Q96DU4 | yes | 14 | CUB (2), ZP (1) | Respiratory and digestive systems. Tears | Epithelial differentiation; innate immune system | T | (56) |
| | Gp340 | Q9UKJ4 | yes | 14 | CUB (2), ZP (1) | Macrophage, respiratory mucosa, small intestine | Binds a collectin family member; opsonin receptor; innate immune response | T/S | (13) (57) |
| | Salivary agglutinin | Q9UKJ4 | yes | 14 | CUB (2), ZP (1) | Saliva | Broad bacterial recognition and agglutination; binds IgA, IgG and other proteins | T/S | (58) |
| | CRP-Ductin (mouse / pig) | Q60997 | yes | 8 | CUB (5), ZP (1) | Gastric, lug mucosa | Epithelial differentiation; innate immune defence; binds bacteria and some proteins | T/S | (59)(60)(61) |
| | Muclin (mouse) | Q9JHU8 | yes | 8 | CUB (5), ZP (1) | Pancreas | Probable Golgi cargo receptor | T | (62) |
| | Vomeroglandin (mouse) | Q9JMJ9 | yes | 7 | CUB (5), ZP (1) | Vomeronasal system | Involved in pheromone perception | | (63) |
| | Ebnerin (rat) | Q62827 | yes | 4 | CUB (3), ZP (1) | Liver, lung, heart, testis, von Ebner's salivary gland | Hepatic tissue repair and differentiation; regulation of taste sensation | T/S | (64)(65) |
| | Pancrin (rat) | Q6QD54 | - | 4 | CUB (3), ZP (1) | Pancreas | - | - | TrEMBL |
| | Hensin (rabbit) | Q95218 | yes | 8 | CUB (2), ZP (1) | Kidney, gastric mucosa | Induction of epithelial differentiation | S | (66)(67) |
| | Mucin (cow) | Q28908 | yes | - | - | Gall bladder epithelium and pancreas | Binds biliary lipids; promotes cholesterol crystal nucleation | S | (68) |

*Hypothetical protein; ‡ Please refer to Table IV.4 for information about these domains

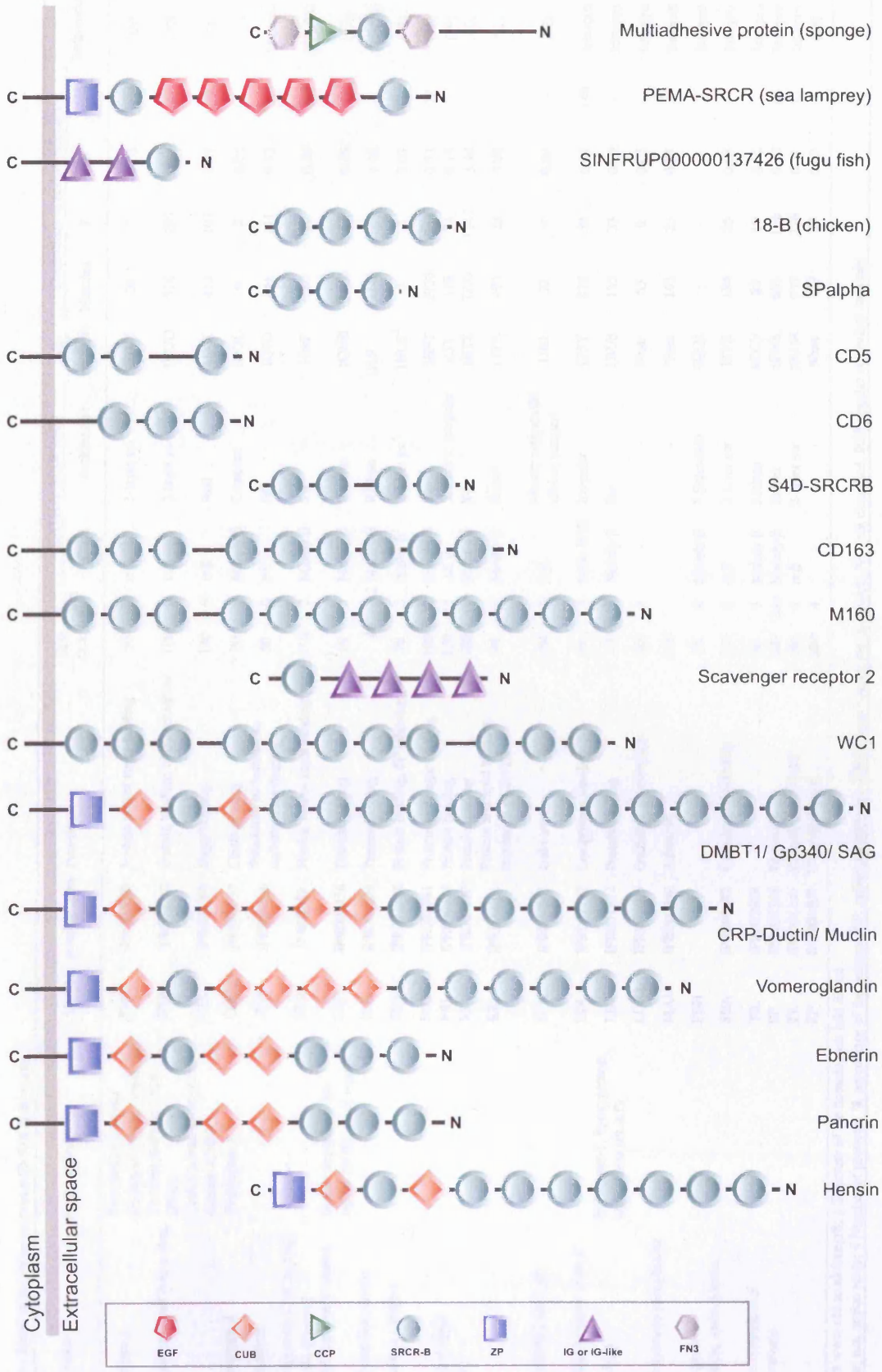


**Figure IV.3** Domain organisation of the protein members of the scavenger receptor cysteine-rich (SRCR) superfamily group B (refer to Tables IV.3 and IV.4 for details and abbreviations).

**Table IV.4 Domains found in the same protein with SRCR domains**

| Name | Other names | Abbr. | Interpro entry | Function | Size (a.a.) | S-S | CATH or SCOP entry | | PDB example | Occurrence in human proteome* | | | | References |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Class | Architecture | | Matches | # | % | Rank | |
| Apple divergent subfamily | Plasminogen, Apple, Nematode (PAN); apple-like | PAN | IPR003609 | Protein and/or sugar binding | 90 | 3 | α β | 3-layer sw | 1BHT | 20 | 9 | 0.03 | - | (69) |
| Broad-Complex, Tramtrack and Bric-a-Brac | Poxvirus and zinc finger (POZ) | BTB | IPR000210 | Protein binding; oligomerisation | 100 | 0 | α β | 2-layer sw | 1BUO | 526 | 209 | 0.63 | 37 | (70) |
| C-type lectin | Carbohydrate-recognition domain (CRD) | CTL | IPR001304 | Sugar binding | 130 | 0 | α β | Roll | 1IXX | 413 | 103 | 0.31 | 79 | (71) |
| Chitin binding domain type 2 | Peritrophin-A | CBD2 | IPR002557 | Chitin binding | 70 | 3 | Mainly β | Complex | 1DQC | 6 | 2 | 0.01 | - | (72) |
| Collagen triple helix repeat | | CLN | IPR008160 | Structural; trimerisation; carbohydrate binding | 50 | 0 | NC | NC | 1Q7D | 861 | 111 | 0.33 | 73 | Interpro |
| Complement subcomponents Clr/Cls, Uegf and Bone morphogenic protein 1 | | CUB | IPR00859 | Protein and/or sugar binding | 110 | 2 | Mainly β | Sw | 1SPP | 680 | 80 | 0.24 | 100 | (73)(74) |
| Domain abundant in complement control proteins | Short complement like repeat (SCR); sushi repeat | CCP | IPR000436 | Protein binding | 60 | 3 | Mainly β | Ribbon | 1QUB | 1420 | 87 | 0.26 | 97 | (75) |
| Epidermal growth factor-like protein | | EGF | IPR006209 | Protein binding | 40 | 2 | Mainly β | Ribbon | 1JL9 | 3740 | 604 | 1.82 | 6 | (75)(76) |
| Factor I membrane attack complex | | FIMAC | IPR003884 | Protein binding; SP inhibitor (?) | 70 | 3 | Alpha β‡ | 2-layer sw‡ | 1BUS‡ | 3 | 2 | 0.01 | - | (77) |
| Fibronectin type III | | FN3 | IPR003961 | Protein and sugar binding | 100 | 2 | Mainly β | Sw | 2HFT | 2930 | 244 | 0.73 | 28 | (78) |
| Frizzled cysteine-rich domain | | FRI | IPR000024 | Protein binding | 120 | 5 | All α | 3 helices; irregular | 1IJY | 118 | 34 | 0.10 | - | (79) |
| Immunoglobulin-like | | IG | IPR007110 | Protein binding | 100 | 1 | Mainly β | Sw | 1ETZ | 7270 | 1142 | 3.44 | 1 | (80) |
| Kringle | | KR | IPR000001 | Protein and lipid binding; regulation of SP/TK | 90 | 3 | Mainly β | Barrel | 1TPK | 493 | 28 | 0.08 | - | (81) |
| Limulus factor C, Coch-5b2 and Lgl1 | | LCCL | IPR004043 | Unknown | 100 | 2 | α β | Mainly antiparallel sheets; unusual | 1JBI | 30 | 9 | 0.03 | - | (82) |
| Low-density lipoprotein receptor class A | | LDL | IPR002172 | Lipoprotein binding | 40 | 3 | Little RSE | Irregular | 1F5Y | 879 | 58 | 0.17 | 149 | Interpro |
| Lypoxygenase homology | Polycystin-1, lipoxygenase, alpha-toxin (PLAT) | LH2 | IPR001024 | Protein binding | 110 | 1 | Mainly β | Sw | 1BU8 | 135 | 33 | 0.10 | - | Interpro |
| Lysyl oxidase-like | | LOXL | IPR001695 | Oxidative deamination | 200 | 4 | - | - | None | 65 | 8 | 0.02 | - | Interpro |
| Meprin, A5, receptor tyrosine phosphatase mu | | MAM | IPR000998 | Adhesion | 170 | | - | - | None | 163 | 23 | 0.07 | - | Interpro |
| Parallel beta-helix repeats | | PBH | - | - | 20 | 0 | Mainly β | 3 Solenoide | 1QQ1 | - | - | - | - | Interpro |
| Sea urchin sperm protein, enterokinase, angrin | | SEA | IPR000082 | Carbohydrate binding | 130 | 0 | α β | 2-layer sw | 1IVZ | 184 | 26 | 0.08 | - | Interpro |
| Trypsin Inhibitor-like cysteine-rich | | TIL | IPR002919 | - | 60 | 5 | Mainly β | Ribbon | 1CCV | 83 | 19 | 0.06 | - | Interpro |
| Trypsin-like serine protease | | SP | IPR001254 | Peptidase | 130 | yes | Mainly β | Barrel | 1FAX | 651 | 166 | 0.50 | 45 | Interpro |
| Tyrosine kinase | | TK | IPR001245 | Phosphotransferase | 100 | 0 | α β | 2-layer sw | 1R1W | 937 | 236 | 0.71 | 33 | Interpro |
| Zona pellucida | | ZP | IPR001507 | Oligomerisation | 260 | 4 | - | - | None | 129 | 23 | 0.07 | - | (83) |

* Integr8 Browser @ www.ebi.ac.uk/integr8; ‡ Structure of the homologous fold Kazal

**S-S,** disulphide bonds; **a.a.** amino acids; **#** Number of proteins; **%** percentage of the genome; **SP**, serine phosphatase; **TK** tyrosine kinase; **sw**, sandwich; **NC**, not classified; **RSE** regular secondary structure

These interactions include the Mac-2 binding protein (M2BP) binding to galectin-3, collagen, fibronectin and nidogen[43], cyclophilin C-associated protein (CyCAP) binding to cyclophilin C[44], and SR-A binding to polyanions[84]. SRCR domains are also involved in cell binding to bacteria, probably via interactions with saccharide moieties on the bacterial surface. Trimeric constructs of the macrophage scavenger receptor MARCO containing only the first 19 residues of its SRCR domain bind bacteria with high affinity. The motifs in the SRCR domain Arg-Gly-Arg and/or Arg-Val-Arg are essential for the binding[85]. A peptide corresponding to residues 11-26 of the N-terminal SRCR domain of gp340, with the motif Gln-Gly-Arg has also been reported to bind bacteria[86]. These RGR and QGR sequence motifs are moderately conserved throughout the SRCR superfamily in the second β-strand of the domain (Figure IV.9).

In this Chapter, I present the first structure of a group B SRCR domain, the doubly substituted (V88D/V97K) N-terminal domain of human CD5 (hereafter referred to as *h*CD5d1m). The sequence identity between *h*CD5d1m and the two previously solved group A domains is lower than 20%. However, the distribution of regular secondary structural elements is shown to be similar in the three structures, but these regions account for less than 35% of the 110 residues of the domain. The remainder of the fold comprises long loops that appear to be mobile in *h*CD5d1m. The presence of the additional C1-C4 disulphide-bond in *h*CD5d1m, which occurs only in group B members of the SRCR superfamily, does not significantly impact in the overall conformation of the molecule. The low sequence identity among members of the superfamily suggests that this fold is a stable scaffold that tolerates great sequence variation in its potentially exposed (long) loop regions, probably allowing substantial variations in ligand specificity across the superfamily.

## Experimental procedures

The experimental work presented in this Chapter was a collective effort of Dr Richard Harris (RH), Dr Diego Esposito (DE) and myself (AG). The initials of the person that performed each part of the experimental work are written after the title of the section. The NMR resonance and assignment data from the glycosylated *h*CD5d1 construct were kindly provided by Dr Mark Pfuhl.

*Sample preparation* (AG)

Samples of V88D/V97K *h*CD5d1 labelled with $^{15}N$ or with $^{13}C$ and $^{15}N$ were prepared as described in previous chapters. Labelling was achieved by growing the cells in supplemented minimal medium using $^{13}C$-glucose and/or $(^{15}NH_4)_2SO_4$ as the sole sources for carbon and nitrogen respectively. The purified protein was dissolved in 90% 50 mM phosphate pH 5.0, 200 mM NaCl, 1mM EDTA, 0.01% (w/v) $NaN_3$ / 10% $D_2O$ at a protein concentration in the region of 1 mM. Both standard Imperial grade 5mm NMR tubes (Wilmad) and susceptibility matched tubes (Shigemi) were used to contain the samples.

*NMR spectroscopy* (RH)

All experiments were recorded on either Varian Unity INOVA or Varian UnityPLUS spectrometers, operating at nominal $^1H$ frequencies of 600 MHz and 500 MHz respectively. Both spectrometers are equipped with a 5mm indirect detection, triple resonance ($^1H/^{13}C/^{15}N$) probe and have a Z-axis pulsed field gradient capability. Chemical shifts were indirectly referenced to sodium 2,2-dimethyl-2-silane-pentane-5-sulphonate (DSS) using absolute frequency ratio to the $^1H$ signals[87].

*Sequential backbone and Cβ resonance assignment at 298 K and 318 K* (AG & RH)

Sequence specific resonance assignments of $H^{15}N$, $^{15}NH$, $^{13}C\alpha$ and also $^{13}C\beta$ nuclei were obtained from the triple resonance correlation spectra HNCA[88], HNCACB[89], and CBCA(CO)NH[90]. The HNCA and HNCACB experiments provide cross-peaks between the Cβ and/or Cα of the current (*i*) and the preceding residues (*i*-1), and the N and NH of residue *i*. The CBCA(CO)NH spectrum correlates the Cα and Cβ of the *i*-1 residue with

the N and NH of residue *i*. Typically, 1024 complex data points were collected in the proton dimension and 64 and 34 complex points in the $^{13}C$ and $^{15}N$ dimension respectively (Please refer to Appendix C for a list of all experimental details). Carbonyl carbons and Hα signals were assigned only at 318K using the 3D HNCO[88] and the HA(CACO)NH[91] experiments, respectively.

*Side chain and NOE resonance assignments at 318 K* (RH)

Resonances of aliphatic side chains were assigned using the $^{15}N$-separated TOCSY-HSQC[92] and the HCCH-TOCSY[93] experiments, each recorded with a mixing time of 20 ms. In an ideal case, the latter spectrum allows the correlation of $^{13}C\alpha$, Hα with all aliphatic side chain protons and carbon nuclei within a given residue, provided there is appropriate H-C-C-H connectivity. The $^{15}N$-separated TOCSY-HSQC correlates side chain protons with the amide proton and amide nitrogen. Proton and carbon resonances within aromatic rings were identified with an 'aromatic' HCCH-TOCSY recorded with a $^{13}C$ mixing time of 12 ms. Interproton NOEs arising from through-space dipolar interactions were identified by recording $^{15}N$ and $^{13}C$-edited NOESY-HSQC spectra with a $^{1}H$ mixing time of 100 ms.

*NMR data processing and analysis* (RH & AG)

All spectra were processed and displayed with the program NMRpipe/NMRDraw[94]. Plot2 from AZARA[95] was used to plot all 1D and 2D NMR spectra for Figures IV.4-7. Spectra were analysed using ANSIG 3.3[96].

A typical processing scheme used to obtain 3D NMR spectra from the free induction decay (FID) data set consisted of the following steps:

$^{1}H$ dimension

1. Subtraction of the water signal from the FID by polynomial fitting by application of a high-pass digital filter.
2. Correlation of the FID with a window function (apodisation) to generate a signal with a smooth decay in order to avoid data truncation artefacts.
3. Doubling the data size by addition of null data to the FID (zero filling) in order to increase digital resolution and signal to noise ratio of the transformed spectrum
4. Application of the Fourier transform to obtain intensity vs. frequency data
5. Manual correction of the phase of the spectrum.

$^{13}C$ and $^{15}N$ dimensions

The interferogram was first extrapolated using linear prediction in order to enhance resolution and then the strategy outlined for the $^{1}H$ dimension was followed, with exception of the water signal subtraction.

*Structure calculation* (DE)

Interproton distance restraints were obtained from the analysis of the 3D $^{1}H$-$^{13}C$ NOESY-HSQC and $^{1}H$-$^{15}N$ NOESY-HSQC spectra. Cross peaks were assigned manually, classified into four categories according to their relative peak intensity, and associated with an interproton distance restraint limit as follows: strong peaks (1.8-2.5 Å), medium peaks (1.8-3.0 Å), weak peaks (1.8-3.5 Å) and very weak peaks (1.8-4.0 Å). For NOEs involving methyl groups, 0.5 Å was added to the upper limit to account for the 3-fold degeneracy and narrow linewidth of these signals. Only unambiguously assigned cross peaks and distance restraints were used in the calculations. Ambiguous distance restraints were filtered out based upon the coordinates of the ensemble in early calculations. Backbone $\phi$ and $\psi$ torsion angle restraints were derived from the pattern of $^{1}H\alpha$, $^{13}C\alpha$, $^{13}C\beta$, $^{13}C'$ and $^{15}NH$ chemical shifts using the database searching program TALOS[97]. Dihedral angle restraints were assigned an error range of $\pm 20^{\circ}$. The presence of the four disulphide bonds, Cys44-Cys86, Cys60-Cys125, Cys81-Cys132 and Cys107-Cys117, were predefined in the topology file for structure calculations, but covalent terms corresponding to these bonds were only turned on after the torsion dynamics step.

Structures of *h*CD5d1m were calculated from the experimental restraints using the CNS-solve program[98] employing the PARALLHDGv5.3 force field parameters[99,100], with the non-bonded energy function of PROLSQ. To induce the proper conformation of prochiral centres (methylene and isopropyl groups), protocols were modified to allow floating stereospecific assignments and active atom swapping[101,102]. The dynamic simulated annealing protocol consisted of two stages: high temperature (10,000 K) torsion angle molecular dynamics for 2200 steps of 45 fs each, and Cartesian dynamics while cooling from 2,000 K to 50 K in 20,000 steps of 5 fs each. After the simulated annealing each structure was energy minimised with 200 steps of Powell minimisation and an ensemble of 100 conformers was calculated. In order to improve the structure quality, all structures generated in the torsion dynamics protocol were used as starting points for restraint

refinement in explicit solvent[100]. The final ensemble consisted of the 20 lowest energy conformers from the refinement.

## Results

*Sequence-specific backbone and Cβ resonance assignment of hCD5d1m at the typical operational temperature of 25°C (298 K)*

The $^{1}H$-$^{15}N$ HSQC spectra of both wild-type *h*CD5d1 and *h*CD5d1m consisted of peaks of different intensities and shapes. Some peaks were strong and symmetric while others were weak and broad; additionally there were fewer signals than expected from the amino acid sequence. This variable peak appearance suggested that a subset of nuclei were experiencing chemical exchange processes on an intermediate timescale, i.e. between $10^{-6}$ to $10^{-3}$ seconds[103]. Such extensive exchange suggests that a substantial proportion of the protein has conformational flexibility. The sequence specific assignment of NH, Cα and Cβ resonances was attempted but the presence of exchange broadening made the assignment of peaks very difficult. Many peaks showed strong overlapping and low intensity, and only 53% of the backbone amide N-H resonances could be unambiguously assigned in the $^{1}H$-$^{15}N$ HSQC spectrum (Figure IV.4).

*The NMR behaviour of hCD5d1m improves with increasing temperature*

The intensity and line width of the NMR signals in the spectra for *h*CD5d1m was greatly affected by changes in temperature (Figures IV.5-6). Very few signals were present in the $^{1}H$-$^{15}N$ HSQC spectrum at 278 K. As the temperature was increased progressively, more peaks were detected and peaks became more intense and symmetric. It was not until 318 K that the intensities of the peaks in the spectrum became fairly even and the vast majority of the expected peaks became visible. This behaviour must be consequence of either of two phenomena: (i) a progressive increase in the rate of the exchange among different conformations; (ii) the stabilisation of a single conformer with the increase in temperature. Whatever the case may be, the *h*CD5d1m samples could be kept at 318 K for several days without denaturation or other changes becoming apparent in the NMR spectra. This high thermal stability, probably due to the presence of multiple disulphide bonds, permitted the full sequence specific backbone resonance assignment of the $^{1}H$-$^{15}N$ HSQC spectrum and

allowed us to proceed with side chain and interproton NOE assignments and finally with the calculation of the 3D solution structure.

*Resonance assignment and calculation of a preliminary structure of hCD5d1m at 318 K*

The backbone amide NH resonances for almost all residues within the sequence of the *h*CD5d1 domain, with the exception of Asn41, Cys44, Ser103 and His126, could be located in the $^{1}H$-$^{15}N$ HSQC spectrum at 318 K (Figure IV.7). The first 10 residues of the construct, which are part of the affinity tag, were not visible in any spectra, indicating that they are highly mobile. The remaining nine residues of the tag were assigned in the $^{1}H$-$^{15}N$ HSQC spectrum. Ninety-five percent of the side chain resonances were assigned. A total of 1088 unambiguous interproton distance restraints and 94 backbone dihedral angle restraints obtained from the analysis of the spectra were used for the structure calculation. The statistics for the 20 conformers with the lowest total energy after refinement in explicit water are presented in Table IV.5 (Figure IV.8A). The average global root mean square deviation (RMSD) of the best-fit ensemble was 1.59±0.26 Å for the backbone and 2.74±0.37 Å for all heavy atoms. The backbone atoms of 61% of the residues (26-30, 33-40, 42, 44-52, 55-64, 76-95, 104-107, 113-114, 117-119, 128-133), including all elements of regular secondary structure and some turns and loops, were better defined within the ensemble and superimposed with an average global backbone RMSD of 1.04±0.14 Å. The positions of the remaining residues were less well defined. These 'disordered' residues are mainly found in three long loops. One loop comprises residues 65 to 75 which correspond to an insert in the amino acid sequence of the CD5d1 domains that is not present in any other SRCR domain (Figure IV.9); the second loop consists of residues 96 to 103; and the last loop connects residues 120 and 128.

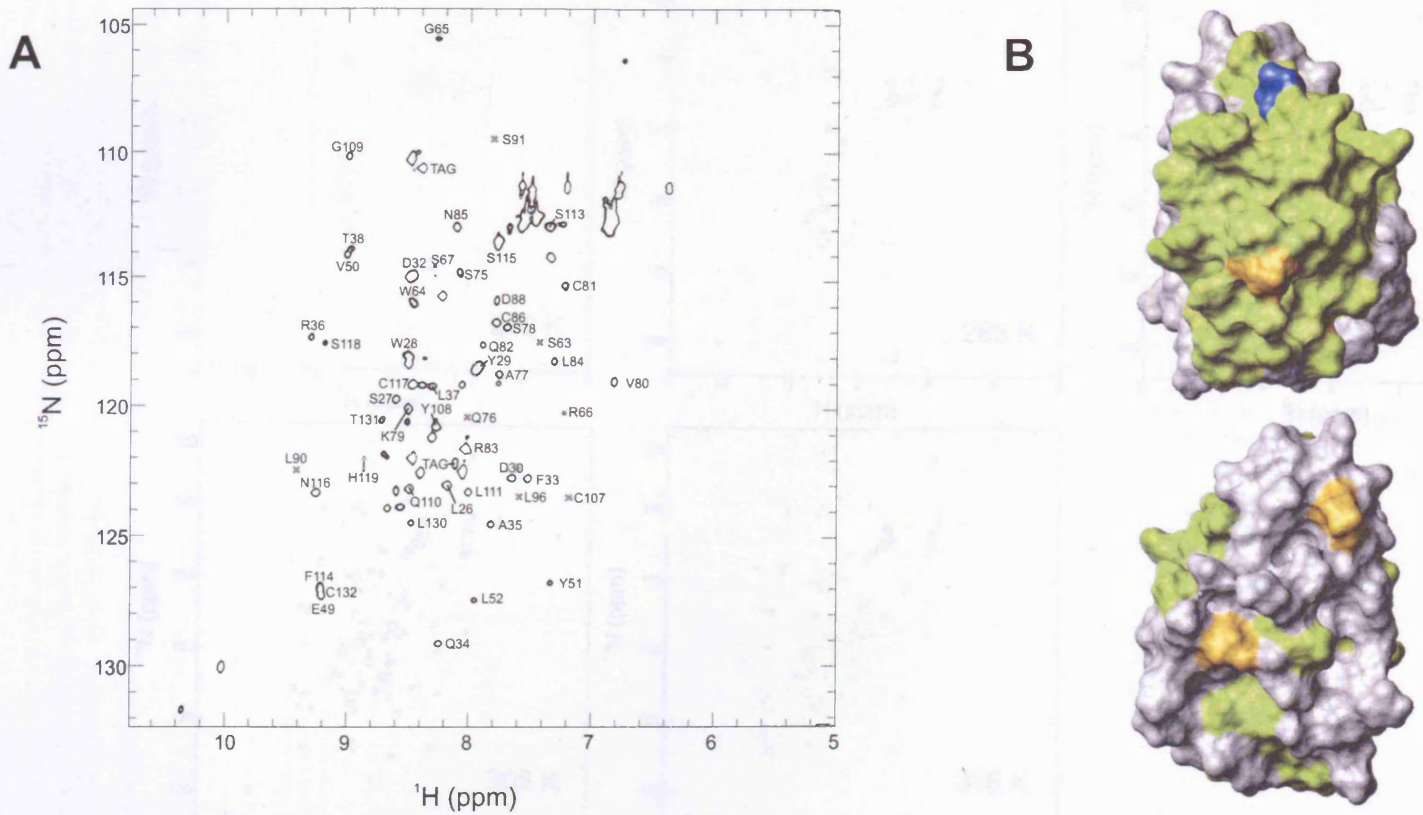


**Figure IV.4** (A) $^{1}$H-$^{15}$N HSQC spectrum of *h*CD5d1m at 298K and pH 6 showing the assignment of the backbone amide resonances. Under these conditions only 40% of the expected backbone HN signals could be assigned. (B) Two views (rotated 180° with respect to each other) of the solvent accessible surface of the final structure of *h*CD5d1m at 318K and pH 5 (see below). Residues whose backbone HN resonance could beassigned at 298K (green) cluster at one side of the N-glycosylation site (blue). Prolines are in yellow.

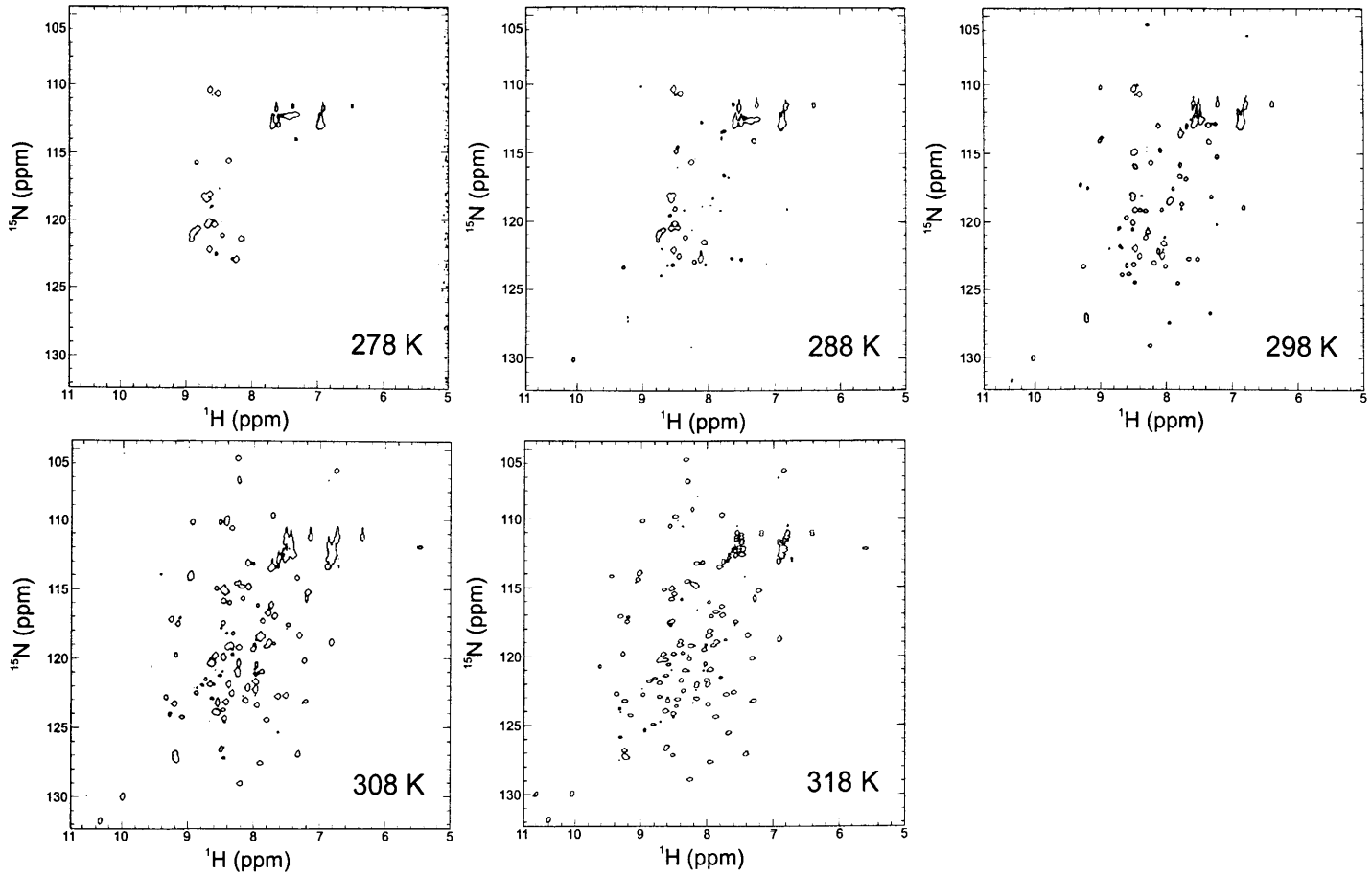


**Figure IV.5** $^{1}$H-$^{15}$N HSQC spectra of *h*CD5d1m at different temperatures. The construct experiences chemical exchange processes, on a timescale between $10^{-6}$ to $10^{-3}$ seconds, that causes severe line broadening and loss of peak intensity. The increase in temperature speeds up the exchange rates to a point where only the averaged conformation is fully represented by backbone NH cross peaks.

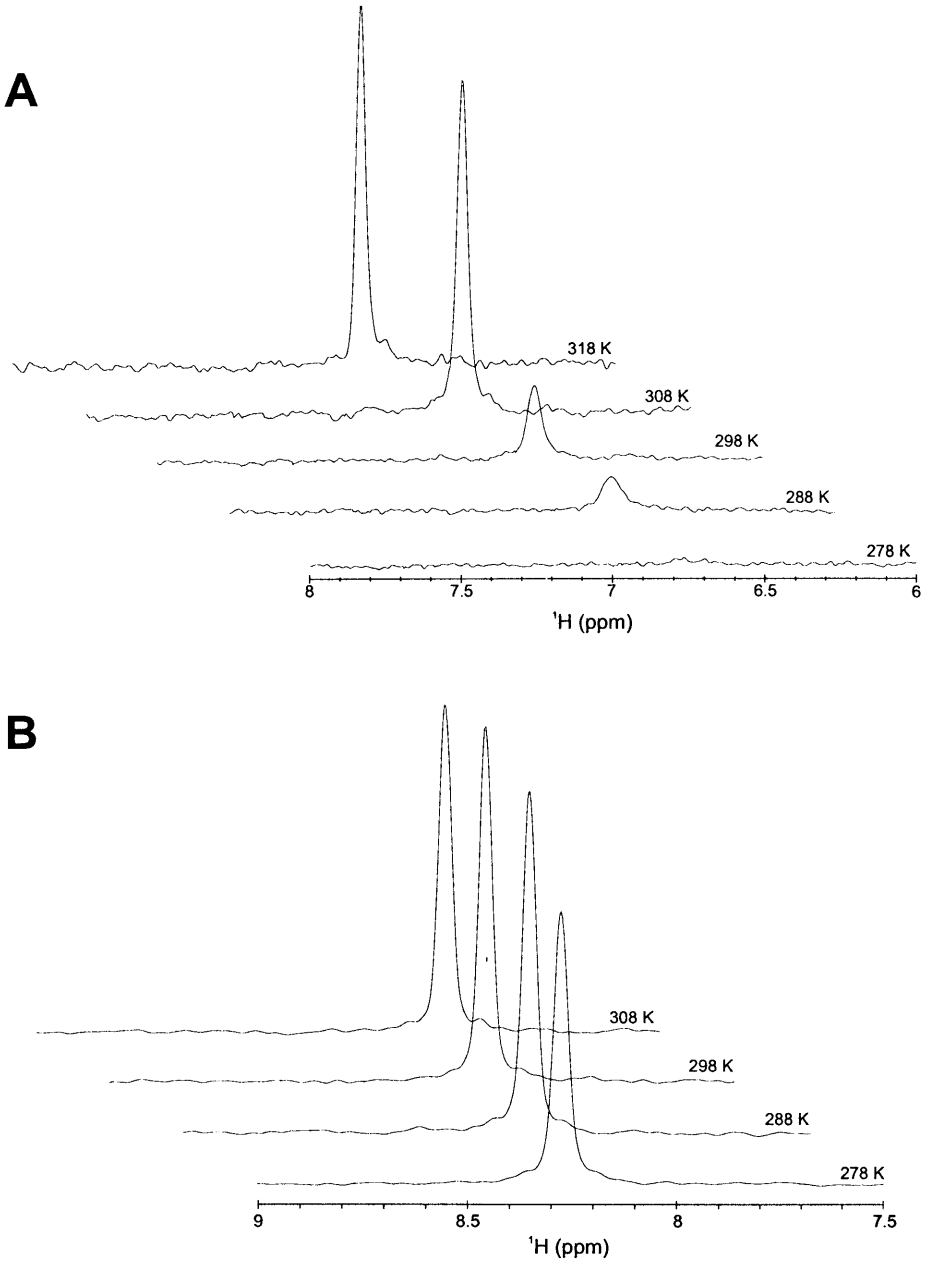


**Figure IV.6** 1D cross section traces through the $^{1}H$-$^{15}N$ HSQC spectra of (A) *h*CD5d1m and (B) human ubiquitin, at several temperatures.
*h*CD5d1m experiences chemical exchange on an intermediate timescale ($10^{-6}$ to $10^{-3}$ second)below 298 K, temperatures above 298 K shift the exchange to a fast timescale ($<10^{-6}$ second) improving the appearance of the spectra. In the case of ubiquitin, temperature has only a minor effect in peak intensity or shape.

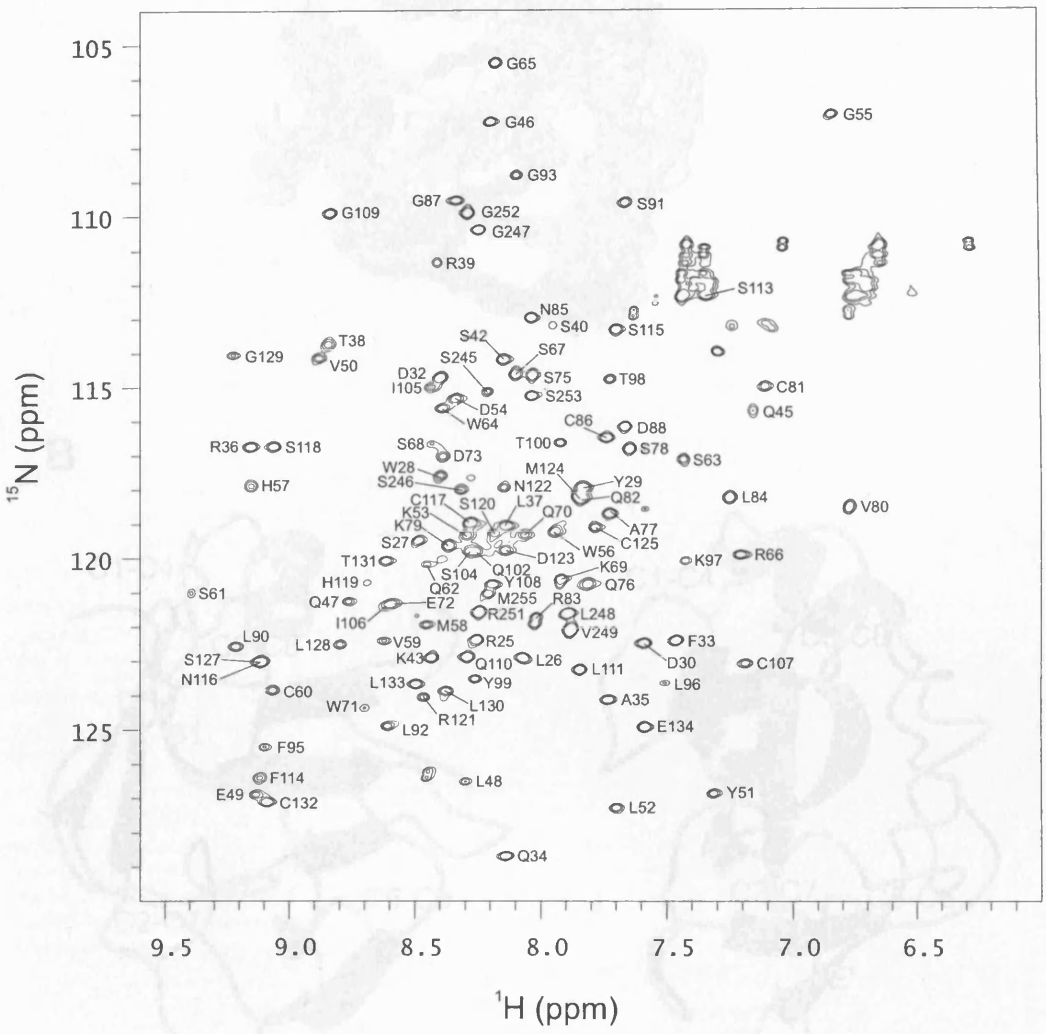


**Figure IV.7** $^{1}$H-$^{15}$N HSQC spectrum of *h*CD5d1m at 318 K and pH 5 showing the assignment of the backbone amide resonaces.

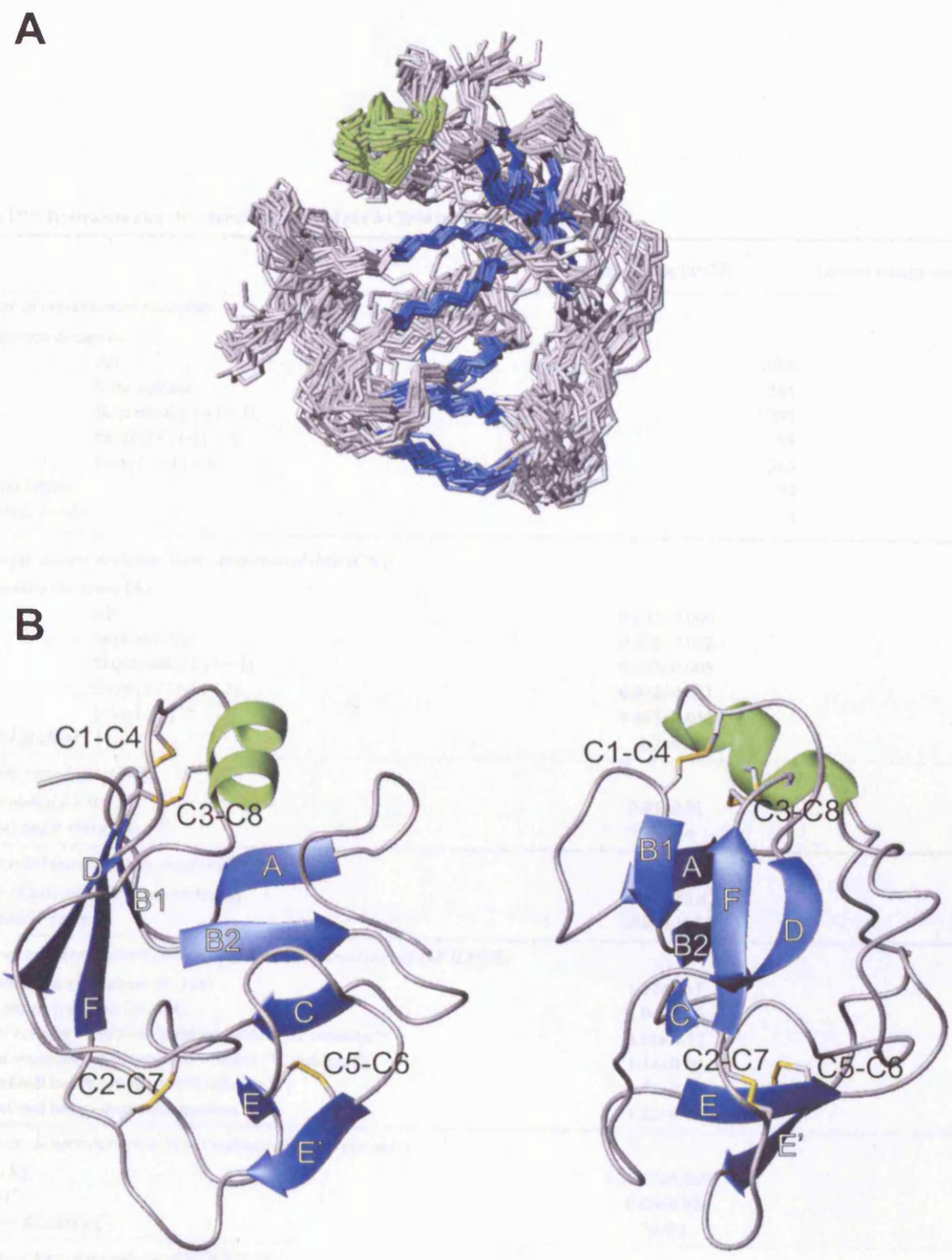


**Figure IV.8** NMR solution structure of *h*CD5d1m at pH 5 and 318 K. (A) Backbone atoms of the ensemble of the 20 lowest energy conformers obtained after restrained refinement in explicit water.
(B) Two orthogonal views of a ribbon representation of the structure of the lowest energy conformer of the ensemble (β-strands, blue; α-helix, green; disulphide bonds, yellow). The β-strands are named according to the number of their corresponding β-sheet (before the point) and β-strand number within that sheet (after the point).

**Table IV.5 Restraints and structural statistics of the *h* CD5d1m ensemble**

| | Ensemble (n=20) | Lowest energy conformer |
|---|---|---|
| *Number of experimental restraints* | | |
| Inter–proton distances | | |
| All | 1088 | |
| Intra–residue | 391 | |
| Sequential (\| i–j \| = 1) | 291 | |
| Short (1< \| i–j \| < 5) | 93 | |
| Long ( \| i–j \| > 5) | 313 | |
| Dihedral angles | 94 | |
| Disulphide bonds | 4 | |
| *Root mean square deviation from experimental data (CNS)* | | |
| Inter–proton distances (Å) | | |
| All | 0.042±0.004 | 0.043 |
| Intra–residue | 0.018±0.002 | 0.018 |
| Sequential (\| i–j \| = 1) | 0.028±0.005 | 0.03 |
| Short (1< \| i–j \| < 5) | 0.072±0.011 | 0.08 |
| Long ( \| i–j \| > 5) | 0.065±0.010 | 0.054 |
| Dihedral angles ( $^{o}$ ) | 1.2±0.1 | 0.89 |
| *Restraint violations (CNS)* | | |
| NOE violations > 0.5 Å | 0.95±0.51 | 1 |
| Dihedral angle violations >5$^{o}$ | 0.2±0.4 | 0 |
| *Non–bonded energy values (kcal/mol) (CNS)* | | |
| van der Waals energy (Lennard-Jones) | -89.0±43.4 | -69.7 |
| Electrostatic energy | -691.7±107.9 | -601.1 |
| *Average root mean square deviation to the mean structure (Å) (MOLMOL)* | | |
| Backbone atoms (residues 25–134) | 1.12±0.17 | 1.2 |
| Heavy atoms (residues 25–134) | 1.94±0.20 | 1.95 |
| Regular econdary structure backbone atoms (32 residues*) | 0.58± 0.12 | 0.74 |
| Regular secondary structure heavy atoms (32 residues*) | 1.14±0.10 | 1.28 |
| Well defined backbone atoms (68 residues**) | 0.71±0.11 | 0.65 |
| Well defined heavy atoms (68 residues**) | 1.22±0.10 | 1.10 |
| *Root mean square deviation from idealised covalent geometry* | | |
| Bonds (Å) | 0.0047±0.0003 | 0.0046 |
| Angles ($^{o}$) | 0.66±0.03 | 0.65 |
| Improper dihedrals ($^{o}$) | 2± 0.2 | 1.9 |
| *Ramachandran plot statistics (PROCHECK)* | | |
| Residues in most favoured regions (%) | 63.2 | 65.3 |
| Residues in additionally allowed regions (%) | 25.3 | 26.3 |
| Residues in generously allowed regions (%) | 5.3 | 2.1 |
| Residues in disallowed regions(%) | 6.3 | 6.3 |

*Residues forming regular secondary structure: 35-37, 45-50, 57-58, 78-84, 91-94, 105-107, 117-119 and 128-131

***Description of the structure***

*h*CD5d1m forms a globular domain of approximately 109 residues, between Arg25 and Glu134. The first residue that is involved in long-range interactions is Leu26, which is seven residues before the nominal N-terminal boundary of the domain deduced from consensus sequence alignment and the two known 3D structures of SRCR domains. These seven residues are not conserved in other species, and possibly this is not a general feature of other group B SRCR domains. The structure consists of a short α-helix, three anti-parallel β-sheets and several extensive loops. As one traverses the polypeptide chain from the N-terminus to the C-terminus, the chain firstly forms four β-strands up to residue Cys60. After a further 19 residues without regular secondary structure the α-helix comes next, followed by a six-residue turn preceding the fifth β-strand. A 10-residue loop follows, then the sixth and the seventh β-strands separated by a further 9-residue loop. After a 6-residue loop, the eighth β-strand is formed at the end of the polypeptide chain. Thus, the regular secondary structural elements of the fold are in sequential order β 1.1 (A)-β2.1 (B1)-β1.2 (B2)-β1.3 (C)-α-β2.2 (D)-β3.1 (E)-β3.2 (E')-β2.3 (F), where the first number indicates the β-sheet number, and the second number indicates the strand number within the β-sheet (Figure IV.8B).

Traversing in the conventional direction from the N-terminus towards the C-terminus, the first disulphide bond in the structure, Cys44 -Cys86, bridges the β 1.1/β2.1 loop to the α/β2.2 loop (this is the C1-C4 disulphide bond that is the distinguishing feature of group B SRCR domains). The second disulphide bond, Cys60-Cys125, corresponding to the generic C2-C7 bond in SRCR domains, connects the β1.3/α and β3.2/β2.3 loops. The disulphide bond between Cys81-Cys132 corresponds to the generic C3-C8 bond, the only disulphide found in all SRCR domains; this disulphide connects the α-helix with the end of strand β2.3, at the C-terminus of the fold. Finally, the disulphide bond between Cys107 and Cys117 (the generic C5-C6 bond) links strand β3.1 with strand β3.2.

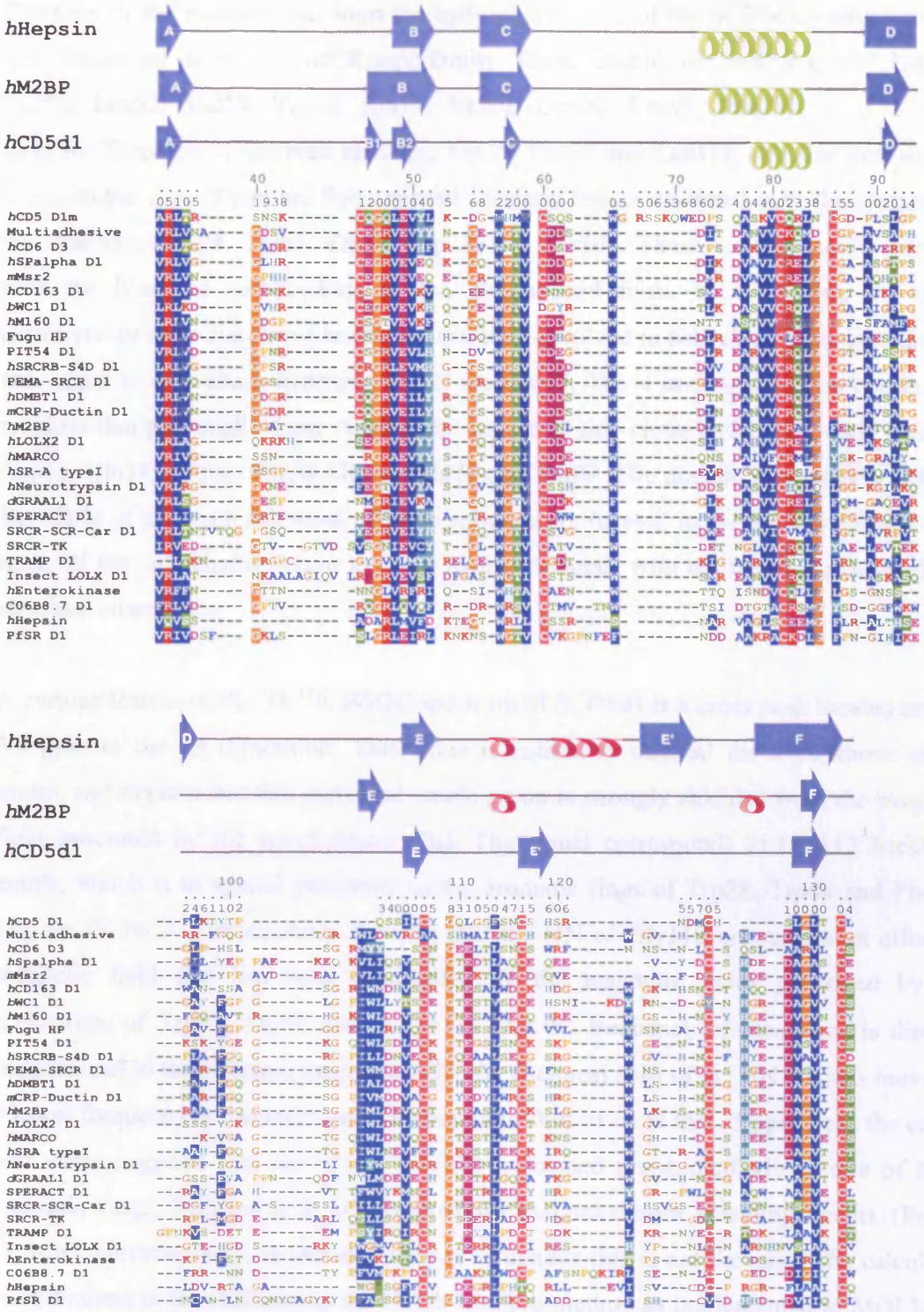


**Figure IV.9** Structure-based multiple sequence aligment of selected members of the SRCR domain superfamily. The regular secondary structure elements of the available structures are represented at the top: β-strands (blue arrows), α-helices (green ribbons) and $3_{1}$ helices (pink ribbons). The alignment was computed with 3D-Coffee using both pairwise sequence aligments and pairwise structure comparisons. Residue colour coding: red, positive; purple, negative; blue, hydrophobic; green, aliphatic & polar; cyan, aromatic & polar; pink, cysteine; yellow, proline; orange, glycine.

Fourteen of the residues that form the hydrophobic core of the *h*CD5d1m structure are well conserved among the SRCR superfamily. These residues are: Ala35, Leu37, Leu48, Val50, Leu52, Val59, Trp64, Ala77, Val80, Leu90, Leu92, Phe114, Leu128, and Leu130. Three non-conserved residues, Trp28, Phe33 and Leu111, are also part of the hydrophobic core. There are five exposed aromatic resides scattered over the surface of the structure: Tyr29, Tyr51, Trp56, Trp71, and Tyr108. These residues could interact with the N-linked carbohydrate in the glycosylated form. In *h*CD5d1m, the only conserved or semi-conserved residues apparently involved in side chain to main chain or side chain to side chain hydrogen bonds are Gln47, Trp64 and Arg83. Non-conserved residues that potentially form structurally important side chain hydrogen bonds include Asp30, Gln34, Arg66, Thr38, Gln62, Gln76 and Thr98. The domain is unusually rich in Ser (16% of all residues); most of these residues are located in the extended loops and many of the side chains appear to make hydrogen bonds with the backbone in the final structure ensemble.

A curious feature of the $^{1}H$-$^{15}N$ HSQC spectrum of *h*CD5d1 is a cross peak located around 3.0 ppm in the $^{1}H$ dimension. This value is extremely unusual for a backbone amide group, and implies that this particular amide group is strongly shielded from the magnetic field generated by the spectrometer ($B_0$). The signal corresponds to Gly112 backbone amide, which is in spatial proximity to the aromatic rings of Trp28, Trp64 and Phe114 (Figure IV.10C). Consequently it seems that the HN of Gly112 encounters an effective magnetic field that has been diminished by the magnetic fields generated by the $\pi$ electrons of Trp28, Trp64 and/or Phe114. As the frequency of resonance is directly proportional to the effective magnetic field, the chemical shift of the HN nucleus moves to a lower frequency in the spectrum. The dramatic shift of more that -5 ppm from the coiled coil value suggests that the HN of Gly112 is located perpendicularly to one of these aromatic rings, because it is at this position that maximum shielding occurs (Figure IV.10B). Unfortunately, in our ensemble of structures this is not the case. The calculated contributions to the shift (using the Haigh-Mallion method as implemented in MOLMOL [104,105]) of the HN nucleus of Gly112 are of roughly -1.0 ppm for Trp28 and -0.5 for Trp64, and the contribution of Phe114 is either null or slightly positive (deshielding) in all members of the ensemble. Most likely, the precise conformation of the aromatic rings or the position of this particular HN is not well defined by the experimental constraints.

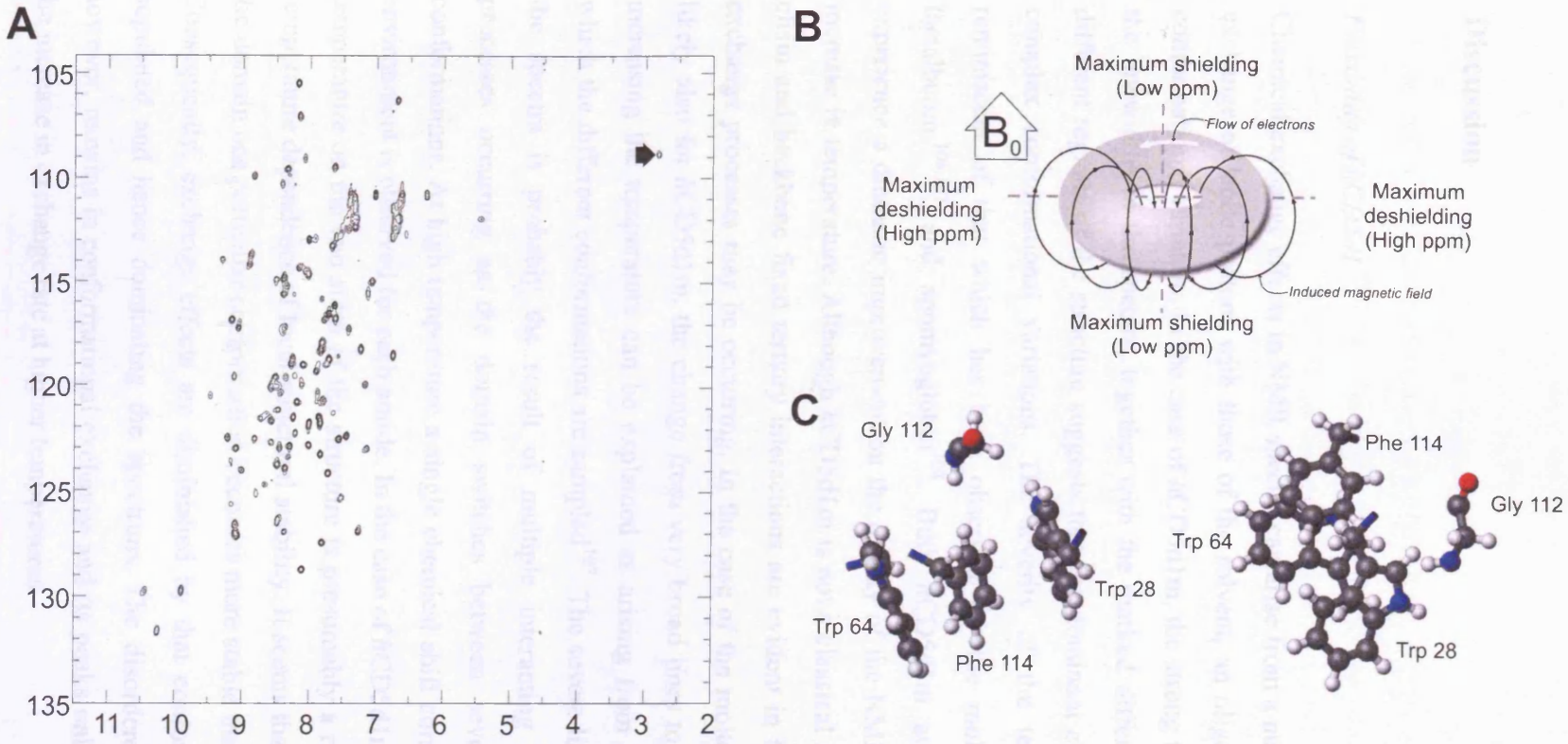


**Figure IV.10** (A) $^{1}H$-$^{15}N$ HSQC spectrum of *h*CD5d1m showing the position of the highly shifted cross peak of the backbone amide group of Gly112. The precession frequency of the HN of Gly 112 is dramatically decreased from the random coil value. Such a large shift usually indicates that the proton is shielded from the external magnetic field by neighbouring aromatic rings, this is called a ring-current shift. As shown in (B), the induced magnetic field generated by the π electrons of an aromatic ring decreases the local field of nuclei situated perpendicular to the ring and increases the local field of nuclei situated in the plane of the ring. Thus, the HN group of Gly112 is predicted to be perpendicular to an aromatic ring. However, the ensemble of calculated structures of *h*CD5d1 does not correspond to this prediction. According to our preliminary ensemble of structures, the aromatic rings that influence the local field of the HN group of Gly112 are Trp28, Trp64 and Phe114 (C). The spatial arrangement of these residues in respect to the HN of Gly112 does not explain the dramatic shift. The ring-current shift for the HN of Gly112 calculated from this structure is only ~1ppm, while the observed shift is of ~ -5.5 ppm

## Discussion

*Flexibility of hCD5d1*

Chemical exchange effects in NMR spectra can arise from a number of sources, including exchange of labile protons with those of the solvent, an oligomerisation equilibrium or conformational variation. In the case of *h*CD5d1m, the strong temperature dependence of the linewidths in the spectra, together with the marked differences in the behaviour of different regions of the structure suggests that the dominant exchange processes involve complex conformational variations. The severity of the temperature dependence is reminiscent of that which has been observed for the molten globule states of α-lactalbumin[106,107] and apomyoglobin[108]. Both *h*CD5d1m and molten globule states experience a dramatic improvement on the quality of the NMR spectra as a result of an increase in temperature. Although *h*CD5d1m is not a classical molten globule (many side chain and backbone fixed tertiary interactions are evident in the NMR data), analogous exchange processes may be occurring. In the case of the molten globule state, and most likely also for *h*CD5d1m, the change from very broad lines to intense narrow lines with increasing the temperature can be explained as arising from an increase in the rate at which the different conformations are sampled[107]. The severe line broadening observed in the spectra is probably the result of multiple interacting conformational exchange processes occurring as the domain switches between several energetically similar conformations. At high temperature a single chemical shift corresponding to its averaged environment is observed for each amide. In the case of *h*CD5d1m, the differential effect of temperature on the two sides of the structure is presumably a result of differences in the temperature dependence of local structural stability. It seems that on the structured face of the domain one particular conformation becomes more stable than the others above 298 K. Consequently, exchange effects are diminished by that conformation becoming highly populated and hence dominating the spectrum. The disordered face of the molecule, however, remains in conformational exchange and its peaks only become visible through the increase in exchange rate at higher temperatures.

It is yet to be determined whether the dynamic features of *h*CD5d1m are an intrinsic property of the domain, or a result of the deliberate experimental manipulations of the protein, *e.g.* the isolation of the domain from the rest of the protein, the mutations and/or the absence of N-linked carbohydrate. The possibility that the two mutations introduced to improve protein solubility are the cause of this behaviour can be discarded, because the NMR spectra of the wild type construct and the double mutant both exhibit the same line broadening. On the other hand, there is strong evidence to support that at least part of this flexibility is due to the absence of the N-linked carbohydrate.

Covalently-bonded glycans play a variety of different roles in proteins. Glycans often provide recognition sites for receptors[109], control of maturation in the endoplasmic reticulum[110], or serve as protein chaperones. It is thought that the glycans in leukocyte co-receptors such as CD5 are involved in preventing non-specific interactions[111], modulating protein organisation within the cell surface[112], protecting the receptors from proteases, providing rigidity to the molecule so that it can extend away from the surface, and increasing conformational stability[112]. Most of the N-H cross peaks in the $^{1}H$-$^{15}N$ HSQC spectrum of an N-linked glycosylated form of CD5d1 expressed in *Pichia pastoris* (Mark Pfhul, personal communication) do not experience line broadening at 298 K, suggesting that the observed mobility of some *h*CD5d1m residues is a consequence of the absence of the N-linked carbohydrate. The fact that above 298 K all of the mobile residues of the unglycosylated domain are clustered on the same face of the fold to one side of the N-glycosylation site (Figure IV.4) supports this explanation. Thus, it appears that the covalently bonded carbohydrate in *h*CD5d1 confers additional rigidity to an extensive region of the structure.

The *P. pastoris* N-glycosylated *h*CD5d1 spectrum exhibits more moderate exchange-derived features that appear to be limited to a group of 10 residues (Ser61, Gln62, Gln70-Glu73, Phe95, and Asp123-Cys125). The cross peaks of these residues are absent in both the N-glycosylated and the unglycosylated $^{1}H$-$^{15}N$ HSQC spectra at 298 K. These residues may have an inherent conformational mobility. The residues Ser61, Gln62, Asp123, Met124, and Cys125 are located close to the disulphide bond between Cys60 and Cys125. It is possible that the motions of these residues are caused by the isomerisation of this disulphide bond. Another disulphide bond that might be experiencing isomerisation is Cys44, whose cross peak could not be found in any of the NMR spectra. A disulphide bond isomerisation resulting from a 180° rotation of the $\chi 1$ dihedral angle of one of the

involved cystines has been comprehensively studied in basic pancreatic trypsin inhibitor (BPTI)[113-115]. In BPTI the interconversion between the two rotamers of the each of the two cystines forming the disulphide bond occurs in a microsecond to millisecond timescale resulting in line broadening of the backbone amide cross peaks of the two cysteines and of their neighbouring (*i*+1) residues[113,116]. Disulphide bond isomerisation has also been reported to contribute to chemical-exchange line broadening in other proteins, such as another Kunitz (BPTI-like) domain[117] and two epidermal growth factor-like domains[118,119].

Another possibility is that the flexibility in the yeast construct arises as a consequence of the non-native nature of the N-linked glycan expressed in *P. pastoris*. The glycoproteins expressed by *P. pastoris* do not contain some of the residues found in mammalian glycans, such as fucose, galactose and sialic acid[120]. These three monosaccharides are known to be components of the N-linked glycans attached to *h*CD5d1 when expressed in mammalian cell lines[121].

Whilst the increase in temperature to 45 °C made possible the calculation of the structure of *h*CD5d1m, it is important to keep in mind that the values of chemical shifts and NOE intensities detected under these conditions in many cases reflect an average over the ensemble of distinct conformations that exist in solution. In order to characterise the dynamic processes involved and make a direct assessment of the amplitude and timescale of these motions, it would be necessary to perform a detailed NMR spin relaxation analysis.

*Comparison with the other SRCR domains*

The only 3D structures of SRCR domains that have been reported to date are the SRCR domain of the tumour associated and matrix protein Mac-2 binding protein (M2BP)[122] (PDB code, 1by2) and the SRCR domain of the serine-protease hepsin[123] (PDB code, 1p57). Both structures were obtained by X-ray diffraction and both belong to group A of the SRCR superfamily, *i.e.* these domains are encoded by two exons and do not have the C1-C4 disulphide bond. At the amino acid sequence level, *h*CD5d1m shares 20% sequence identity and 33% similarity with M2BP SRCR, and 14% identity and 25% similarity with hepsin SRCR. Similarly to *h*CD5d1m, the group A SRCR domains consist of long loops, an α-helix and several β-strands (Figure IV.11). Overall, the group A

structures are more similar to each other (backbone atom RMSD 0.95 Å) than to *h*CD5d1m, although *h*CD5d1m differs less from the M2BP SRCR domain (backbone atom RMSD 1.52 Å) than from the hepsin SRCR domain (backbone atom RMSD of 1.83 Å). In the first 25 residues of the fold, which includes both elements of regular secondary structure and loops, the structure of M2BP SRCR is categorically more similar to *h*CD5d1m than to hepsin SRCR (backbone atom RMSD 1.24 Å and 7.14 Å, respectively), and this could have some functional significance.

The architecture of the SRCR domains is unusually subtle and irregular, and the automatic analysis of the topology of the three structures gives conflicting results. The Kabsch and Sander (DSSP) algorithm, defines secondary structure elements based on hydrogen bond patterns and it is the most widely used method to assign secondary structure[124]. The algorithm attributes a different organisation to the β-strands in each of the group A SRCR domain structures: A single six-stranded β-sheet in M2BP SRCR and seven β-strands forming two β-sheets in the case of the hepsin SRCR. The M2BP SRCR domain was the first of the structures to be solved and the six β-strands were denoted A through F (Figure IV.9); the same nomenclature will be adopted hereafter for consistency. For the region of the structure corresponding to strand B in the M2BP SRCR domain, DSSP assigns two separate β-strands in *h*CD5d1, here labelled B1 and B2. An additional strand found in the hepsin SRCR domain and *h*CD5d1m, here referred to as E', runs between strands E and F. In the hepsin SRCR domain, strands A, B, C, E and E' are assigned to one β-sheet and strands D and F to the other. In *h*CD5d1m, strands A, B2 and C are assigned to one β-sheet, strands B1, D and F to a second, and strands E and E' to a third sheet. Strands A-D and F, that are supposedly part of different elements of secondary structure in each case, are the best conserved among the three structures and superimpose with an average global backbone atom RMSD of 1.43±0.45 Å. Several other regions of the structure superimpose well if considered independently from the rest of the structure (Table IV.6), which suggests that some parts of the fold form rigid structural motifs with hinges between them.

Despite the variability in automatically attributed β-sheet organisation, the tertiary contacts established in the three structures are very similar, as can be seen in their Cα-Cα contact maps (Figure IV.13). There are contacts in all the three structures linking strands A to B, B to C, B to F, C to E, C to F, D to F. There are also contacts between the sequence segment corresponding to strands E and E' in the three structures, regardless of the fact that the M2BP SRCR domain lacks a specifically defined E' strand. This connectivity pattern suggests that all the three structures form a very irregular curved six or seven-stranded mixed β-sheet, similar to that originally described for the M2BP SRCR domain.

Prior to the solution of the *h*CD5d1m 3D structure, it was expected that the C1-C4 disulphide would play a major role in shaping the fold, because it is the only consistent difference –at the amino acid sequence level– between Group A and Group B SRCR domains. However this does not turn out to be the case. At this point is not evident from the *h*CD5d1m structure what the significance of this disulphide bond is to the overall properties of the fold. Cys44 (C1) of *h*CD5d1m is located in the AB loop. This loop has a very similar conformation in both M2BP SRCR and *h*CD5d1m (backbone atom RMSD 0.88 Å), implying that its shape is not a consequence of the presence of the disulphide. However, it does appear that the C1-C4 disulphide bond is reinforcing C3-C8 in its role to maintain the α-helix in position with respect to strands F and D. In the group A structures, the α-helix lies between strands D and F on one side and the EE' loop on the other. In the *h*CD5d1m structure the α-helix also packs against the D and F strands, but it does not interact with the EE' loop. Instead, side chains from the α-helix establish hydrophobic interactions with the N-terminal region, prior to strand A, which is not a part of the fold in the group A structures. As was noticed above, this unexpected contribution of the N-terminus to the structure is not likely to be conserved in other group B SRCR domains, and its functional significance, if any, remains to be established.

The only disulphide bond that maintains a consistent position in the three SRCR domain structures is C3-C8, which joins the α-helix with the domain C-terminus. This is the single disulphide bond shared by all SRCR domains, and presumably it is essential for the formation of the overall fold. The M2BP SRCR domain and the hepsin SRCR domain share the position of the C2-C7 disulphide, and the equivalent disulphide in *h*CD5d1m is only 4 Å away. The C5-C6 disulphide is in a similar position in *h*CD5d1m and the hepsin

SRCR domain, but the C5-C6 disulphide in M2BP SRCR domain is 13 Å away. This difference was quite unanticipated, because it is intuitively expected that the conservation of the pattern of cysteines at the amino acid sequence level would be reflected in a similar distribution of the disulphide bonds in space.

A very interesting feature of the hepsin SRCR 3D structure is that it was solved together with the hepsin serine-protease domain, giving the first glimpse of the potential, quaternary organisation of SRCR domains. The interdomain interface lies roughly perpendicular to the SRCR domain α-helix and corresponds to the area formed by the C-terminus of the α-helix and the αD turn the middle part of the EE' loop, and both the N and C-termini of the domain. The most prominent residues in the contact area are those of the αD turn, whose positions are very well conserved in the three SRCR structures (backbone atom RMSD 0.4 ± 0.16 Å). This region is also where the C1-C4 disulphide bond is located in the *h*CD5d1m structure. Perhaps this zone is optimised to be the interface with different domains in SRCR group A, while in group B the presence of the disulphide and the significantly shorter EE' loop facilitate the stacking of several SRCR domains. This would explain why group B SRCR domains form tandem arrays whilst the Chordate group A SRCR domains are mostly found in mosaic proteins.

*Ligand binding*

The only detailed protein-protein interaction studies involving an SRCR domain are those of the C-terminal SRCR domain of CD6 (CD6d3) with the N-terminal Ig C-like domain of CD166[125-127]. Mutagenesis and modelling studies have established that the binding site of CD6d3 comprises residues Glu293, Tyr327, Ser329, Phe344, Arg346, Arg348, Gln352 and Ser353[128]. According to sequence alignment with the known SRCR 3D structures (Figure IV.9), most of these residues are in the E'F loop of the fold, except for Glu293 which is in the Cα loop and Tyr327 and Ser329 which are at the beginning of the EE' loop. The conservation of the loop length and amino acid composition of the EE' and E'F loops is very poor among members of the SRCR superfamily, so perhaps variability in these loops modifies the ligand specificity. These two loops lie on the opposite side of the hepsin SRCR domain interface with its correspondent serine protease domain, lending support to the hypothesis that these two separate faces of the domain could be consensus ligand binding and intra-protein interaction zones.

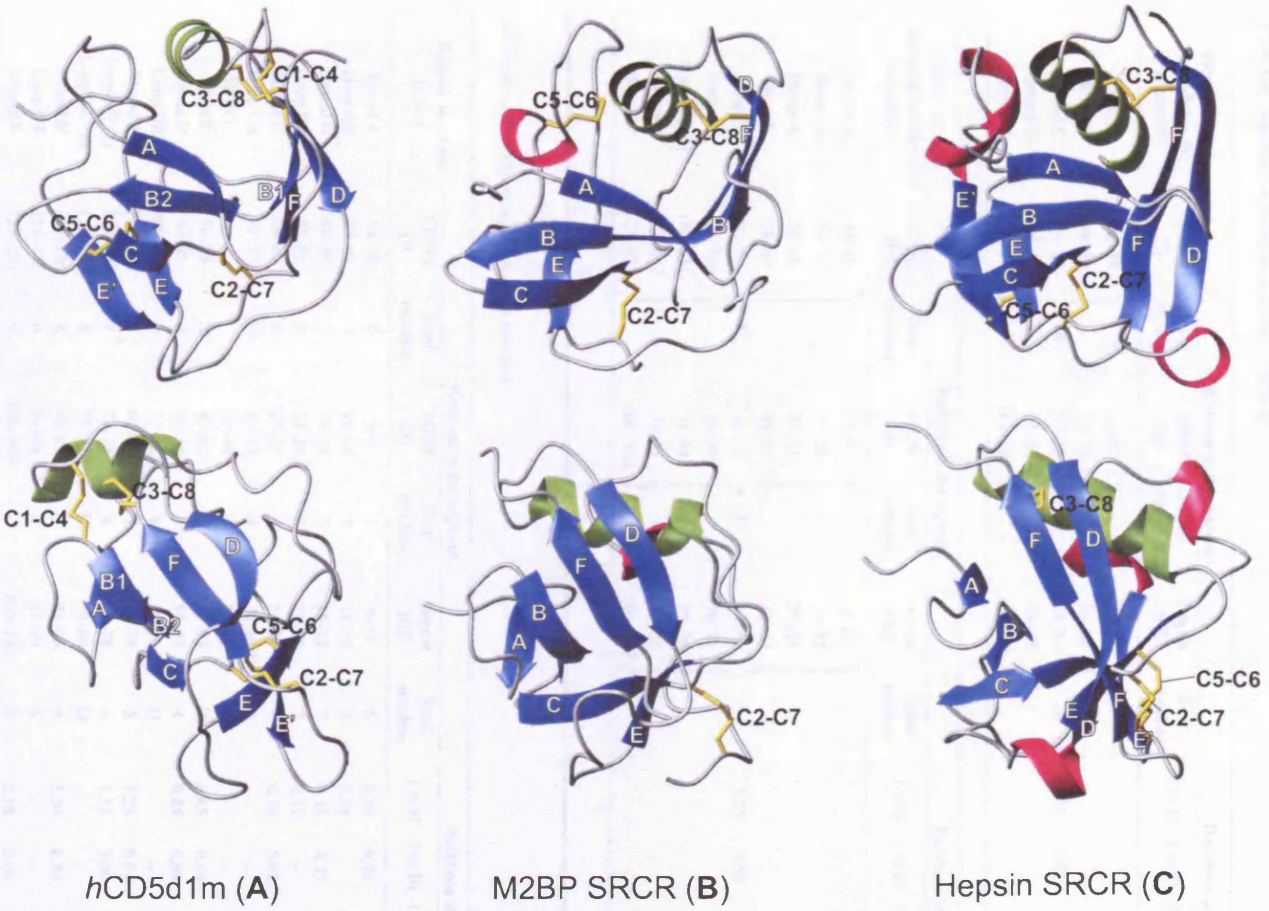


**Figure IV.11** Orthogonal views of a ribbon representation of the three known structures of scavenger receptor cysteine-rich (SRCR) domains. The β-strands (in blue) are labelled sequentially A-B-C-(α helix)-D-E-E'-F; α-helices are in green and $3_{10}$ helices are in pink.
(A) The NMR solution structure of *h*CD5d1m
(B) The crystal structure of the SRCR domain of the tumor-associated antigen and matrix protein Mac-2 binding protein (PDB code: 1by2)
(C) The crystal structure of the SRCR domain of the extracellular serine protease hepsin (PDB code: 1p57)

**Table IV.6 Structural alignments of the SRCR domains**

*Rigid body alignments of homologous tertiary structures*

| Structure feature included in the alignment | Residues in the segment | | | | | | Backbone atom RMSD (Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CD5d1 (I) | Total residues | M2BP (II) | Total residues | Hepsin (III) | Total residues | I vs II | I vs III | II vs III | Average |
| Strand A | 34–38 | | 5–9 | | 8–12 | | | | | |
| Strand B | 46–52 | | 17–23 | | 16–22 | | | | | |
| Strand C | 56–59 | 24 | 27–30 | 24 | 28–31 | 24 | 1.52 | 1.83 | 0.95 | 1.43±0.45 |
| Strand D | 91–93 | | 54–56 | | 55–57 | | | | | |
| Strand F | 128–132 | | 101–105 | | 101–105 | | | | | |

| Related structure feature included in the alignment | Residues in the segment | | | | | | Backbone atom RMSD (Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CD5d1 (I) | Total residues | M2BP (II) | Total residues | Hepsin (III) | Total residues | I vs II | I vs III | II vs III | Averge |
| Strand A | 34–38 | | 5–9 | | 8–12 | | | | | |
| Strand B | 46–52 | | 17–23 | | 16–22 | | | | | |
| Strand C | 56–60 | | 27–31 | | 28–32 | | | | | |
| Helix2 | 78–89 | | 41–52 | | 42–53 | | | | | |
| Strand D | 90–93 | 51 | 53–56 | 51 | 54–57 | 51 | 3.25 | 3.50 | 1.79 | 2.85±0.92 |
| Strand E | 104–108 | | 66–70 | | 70–74 | | | | | |
| Loop EE' (1) | 109–112 | | 71–74 | | 75–78 | | | | | |
| Strand E' | 117–120 | | 92–95 | | 93–96 | | | | | |
| Strand F | 127–132 | | 100–105 | | 100–105 | | | | | |

*Alignment of homologous segments of the structure*

| Related structure feature | Residues in the segment | | | | | | Backbone atom RMSD (Å) | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CD5d1 (I) | Total residues | M2BP (II) | Total residues | Hepsin (III) | Total residues | I vs II | I vs III | II vs III | Average |
| Strand A | 34–38 | 5 | 5–9 | 5 | 8–12 | 5 | 0.59 | 0.93 | 0.83 | 0.78±0.18 |
| Loop AB | 39–45 | 7 | 10–16 | 7 | 13–15 | 3 | 0.88 | – | – | – |
| Strand B | 46–52 | 7 | 17–23 | 7 | 16–22 | 7 | 0.33 | 0.73 | 0.67 | 0.58±0.21 |
| Loop BC | 53–55 | 3 | 24–26 | 3 | 23–27 | 7 | 0.71 | – | – | – |
| Strand C | 56–60 | 4 | 27–31 | 4 | 28–32 | 4 | 0.76 | 0.68 | 0.40 | 0.61±0.19 |
| Loop Cα | 61–77 | 15 | 32–35 | 4 | 33,34,36,37 | 4 | – | – | 1.08 | – |
| Helix1 | – | | 37–40 | 4 | 38–41 | 5 | – | – | 0.10 | – |
| Helix2 | 78–89 | 12 | 41–52 | 12 | 42–53 | 12 | 0.65 | 0.64 | 0.22 | 0.51±0.25 |
| Strand D | 90–93 | 4 | 53–56 | 4 | 54–57 | 4 | 0.84 | 0.98 | 0.52 | 0.98±0.24 |
| Loop DE | 94–103 | 9 | 57–65 | 8 | 58–69 | 11 | – | – | – | – |
| Strand E | 104–108 | 5 | 66–70 | 5 | 70–74 | 5 | 1.23 | 0.94 | 1.08 | 1.08±0.13 |
| Loop EE' (1) | 109–112 | 4 | 71–74 | 4 | 75–78 | 4 | 1.11 | 1.00 | 1.12 | 1.08±0.06 |
| Loop EE' (2) | 113–116 | 4 | 75–91 | 17 | 79–92 | 14 | – | – | – | – |
| Strand E' | 117–120 | 4 | 92–95 | 4 | 93–96 | 4 | 1.04 | 0.89 | 1.28 | 1.07±0.20 |
| Loop E'F | 121–126 | 6 | 96–99 | 4 | 97–99 | 3 | – | – | – | – |
| Strand F | 127–132 | 6 | 100–105 | 6 | 100–105 | 6 | 0.59 | 0.66 | 0.48 | 0.55±0.10 |

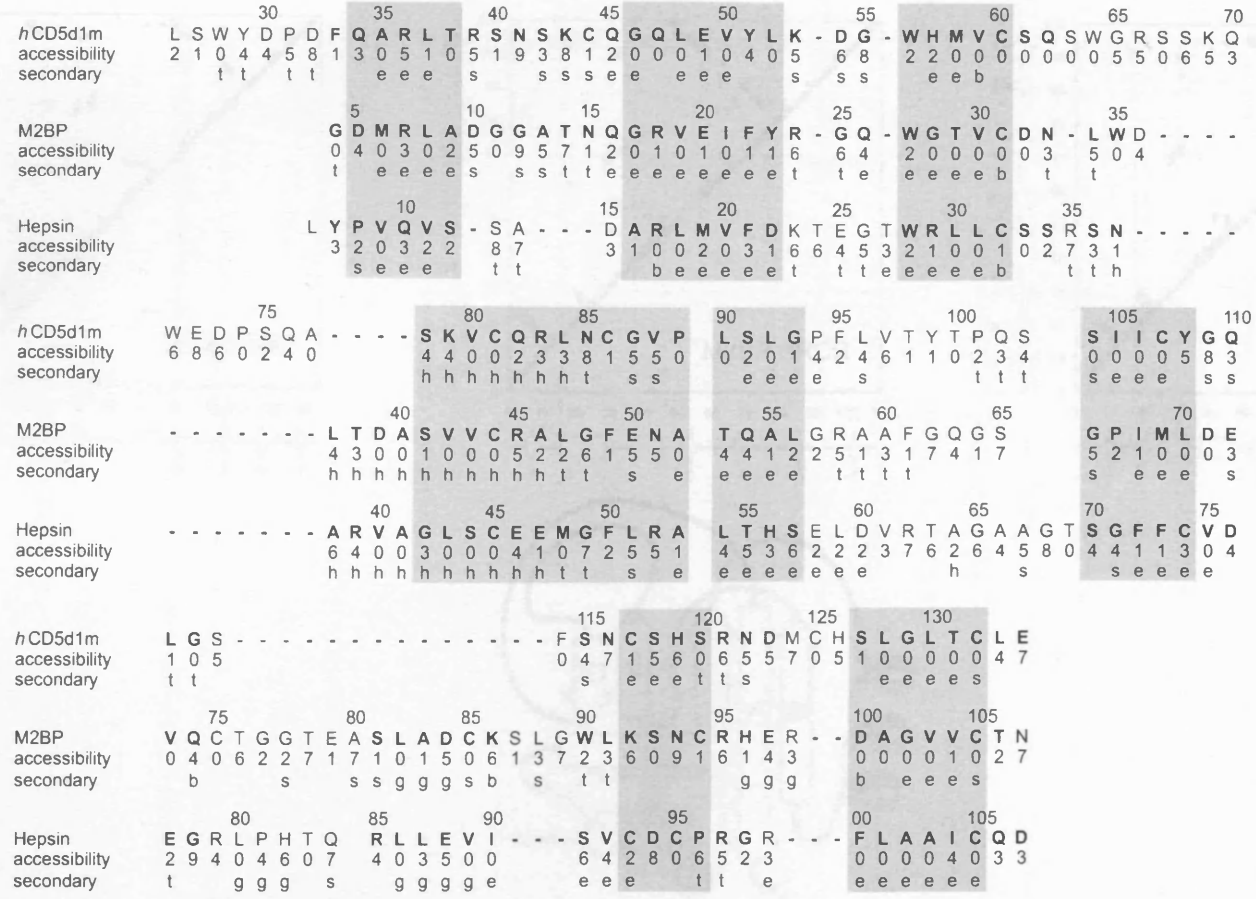

**Figure IV.12** Structure-based amino acid sequence alignment of *h*CD5d1m, M2BP and hepsin SRCR domains. Shaded regions indicate the more highly conserved segments of the structures. Residues are shown in bold when there is an acceptable degree of structural similarity. The secondary structure assignment is from DSSP: α-helix (h), $3_{10}$helix (g), helix-turn (t), extended β-sheet (e), β-bridge (b) and bend (s). Side chain solvent accessibility values were computed with NACCESS and then converted to a scale from 0 (0-10% relative solvent accesibility) to 9 (90-100% relative solvent accessibility).

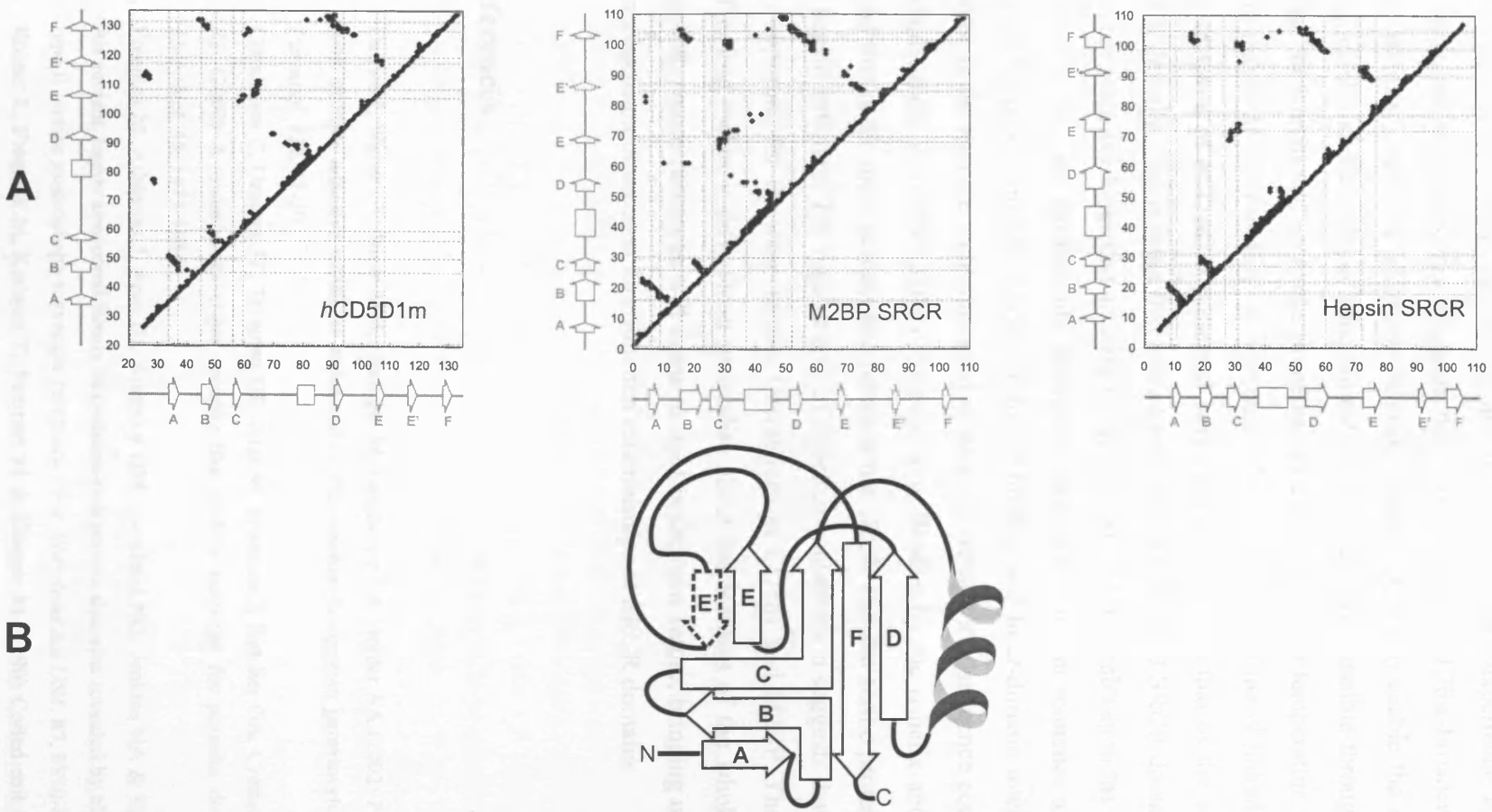


**Figure IV.13** (A) Cα-Cα contact map (distance between atoms < 6 Å) for the three known SRCR domain structures. The nature of the contacts among the strands is very similar in the three structures. There are antiparallel contacts (bands perpendicular to the main diagonal) between β-strands (or sequence elements equivalent to) A and B, B and C, C and F, D and F, and E and E' (Note that the E' strand is not properly defined in M2BP SRCR domain structure). The contacts between strands C and E are of a parallel nature (band parallel to the main diagonal). (B) Schematic illustration of the common topology of the three known SRCR domains, representing the interaction among the β-strands derived from the contact maps. The three structures form a highly distorted six (or seven) strand mixed β-sheet.

In this chapter, I have presented the results of the NMR study of an unglycosylated double residue mutant of the N-terminal domain of CD5d1 (*h*CD5d1m) and discussed their implications. Some regions of this protein construct experience mobility in the microsecond to millisecond timescale that causes substantial line-broadening and cross peak intensity loss. The recording of good quality spectra to enable the complete peak assignment and 3D solution structure calculation was only possible through acceleration of the conformational exchange processes by elevation of the temperature to 318 K. The main reason of this mobility is most probably the absence of the N-linked carbohydrate. The structure of *h*CD5d1m presented here closely resembles that of the other published SRCR domains, and it is the first solved structure of a group B SRCR domain. The SRCR fold has extensive loops that account for at least 50% of the residues in the domain. Some of these loops vary significantly among the structures, both in sequence and in position, and are the most probable regions for ligand binding and inter-domain interactions. Other regions in the fold are well conserved in structure despite low sequence conservation. For example, there is a structurally conserved turn, flanked by the α-helix and the fourth β-strand that is the main contact site between the SRCR and the serine protease domains in the hepsin structure. The high degree of structure conservation suggests that this feature is also important for the inter-domain interactions of CD5d1 and M2BP. The nature of the SRCR fold makes it difficult to generalise about the features of the whole superfamily. Hopefully, more structures will come to light in the near future, bringing us closer to full comprehension of the structure-function relationships of SRCR domains.

## References


1. Delrieu I, Waller CC, Mota MM, Grainger M, Langhorne J & Holder AA (2002) PSLAP, a protein with multiple adhesive motifs, is expressed in *Plasmodium falciparum* gametocytes. *Mol Biochem Parasitol.* **121**, 11-20.
2. Claudianos C, Dessens JT, Trueman HE, Arai M, Mendoza J, Butcher GA, Crompton T & Sinden RE (2002) A malaria scavenger receptor-like protein essential for parasite development. *Mol Microbiol.* **45**, 1473-1484.
3. Freeman M, Ashkenas J, Rees DJ, Kingsley DM, Copeland NG, Jenkins NA & Krieger M (1990) An ancient, highly conserved family of cysteine-rich protein domains revealed by cloning type I and type II murine macrophage scavenger receptors. *Proc Natl Acad Sci USA.* **87**, 8810-8814.
4. Rohrer L, Freeman M, Kodama T, Penman M & Krieger M (1990) Coiled-coil fibrous domains mediate ligand binding by macrophage scavenger receptor type II. *Nature.* **343**, 570-572.

5. Resnick D, Chatterton JE, Schwartz K, Slayter H & Krieger M (1996) Structures of class A macrophage scavenger receptors. Electron microscopic study of flexible, multidomain, fibrous proteins and determination of the disulfide bond pattern of the scavenger receptor cysteine-rich domain. *J Biol Chem.* **271**, 26924-26930.
6. McAlister MS, Brown MH, Willis AC, Rudd PM, Harvey DJ, Aplin R, Shotton DM, Dwek RA, Barclay AN & Driscoll PC (1998) Structural analysis of the CD5 antigen--expression, disulphide bond analysis and physical characterisation of CD5 scavenger receptor superfamily domain 1. *Eur J Biochem.* **257**, 131-141.
7. Hohenester E, Sasaki T & Timpl R (1999) Crystal structure of a scavenger receptor cysteine-rich domain sheds light on an ancient superfamily. *Nat Struct Biol.* **6**, 228-232.
8. Somoza JR, Ho JD, Luong C, Ghate M, Sprengeler PA, Mortara K, Shrader WD, Sperandio D, Chan H, McGrath ME & Katz BA (2003) The structure of the extracellular region of human hepsin reveals a serine protease domain and a novel scavenger receptor cysteine-rich (SRCR) domain. *Structure.* **11**, 1123-1131.
9. Resnick D, Pearson A & Krieger M (1994) The SRCR superfamily: a family reminiscent of the Ig superfamily. *Trends Biochem Sci.* **19**, 5-8.
10. Aruffo A, Bowen MA, Patel DD, Haynes BF, Starling GC, Gebe JA & Bajorath J (1997) CD6-ligand interactions: a paradigm for SRCR domain function? *Immunol Today.* **18**, 498-504.
11. Graversen JH, Madsen M & Moestrup SK (2002) CD163: a signal receptor scavenging haptoglobin-hemoglobin complexes from plasma. *Int J Biochem Cell Biol.* **34**, 309-314.
12. Tino MJ & Wright JR (1999) Glycoprotein-340 binds surfactant protein-A (SP-A) and stimulates alveolar macrophage migration in an SP-A-independent manner. *Am J Resp Cell Mol.* **20**, 759-768.
13. Holmskov U, Mollenhauer J, Madsen J, Vitved L, Gronlund J, Tornoe I, Kliem A, Reid KB, Poustka A & Skjodt K (1999) Cloning of gp-340, a putative opsonin receptor for lung surfactant protein D. *Proc Natl Acad Sci USA.* **96**, 10794-10799.
14. Wu ZW, Golub E, Abrams WR & Malamud D (2004) gp340 (SAG) binds to the V3 sequence of gp120 important for chemokine receptor interaction. *AIDS Res Hum Retrov.* **20**, 600-607.
15. Blumbach B, Pancer Z, Diehl-Seifert B, Steffen R, Munkner J, Muller I & Muller WE (1998) The putative sponge aggregation receptor. Isolation and characterization of a molecule composed of scavenger receptor cysteine-rich domains and short consensus repeats. *J Cell Sci.* **111**, 2635-2644.
16. Pancer Z, Munkner J, Muller I & Muller WE (1997) A novel member of an ancient superfamily: sponge (*Geodia cydonium*, Porifera) putative protein that features scavenger receptor cysteine-rich repeats. *Gene.* **193**, 211-218.
17. Krasko A, Schroder HC, Batel R, Grebenjuk VA, Steffen R, Muller IM & Muller WEG (2002) Iron induces proliferation and morphogenesis in primmorphs from the marine sponge *Suberites domuncula. DNA Cell Biol.* **21**, 67-80.
18. Dangott LJ, Jordan JE, Bellet RA & Garbers DL (1989) Cloning of the mRNA for the protein that crosslinks to the egg peptide speract. *Proc Natl Acad Sci USA.* **86**, 2128-2132.
19. Shimizu T, Yoshino K & Suzuki N (1994) Identification and characterisation of putative receptors for Sperm-activating peptide I (SAP-I) in spermatozoa of the sea urchin *Hemicentrotus pulcherrimus. Dev Growth Differ.* **36**, 209-221.

20. Pancer Z, Rast JP & Davidson EH (1999) Origins of immunity: transcription factors and homologues of effector genes of the vertebrate immune system expressed in sea urchin coelomocytes. *Immunogenetics*. **49**, 773-786.
21. Pancer Z (2000) Dynamic expression of multiple scavenger receptor cysteine-rich genes in coelomocytes of the purple sea urchin. *Proc Natl Acad Sci USA*. **97**, 13156-13161.
22. Gorman MJ, Andreeva OV & Paskewitz SM (2000) Sp22D: a multidomain serine protease with a putative role in insect immunity. *Gene*. **251**, 9-17.
23. Irving P, Troxler L, Heuer TS, Belvin M, Kopczynski C, Reichhart JM, Hoffmann JA & Hetru C (2001) A genome-wide analysis of immune responses in *Drosophila*. *Proc Natl Acad Sci USA*. **98**, 15119-15124.
24. Molnar J, Fong KSK, He QP, Hayashi K, Kim Y, Fong SFT, Fogelgren B, Szauter KM, Mink A & Csiszar K (2003) Structural and functional diversity of lysyl oxidase and the LOX-like proteins. *BBA-Proteins Proteom*. **1647**, 220-224.
25. Ohashi M, Kawamura K, Fujii N, Yubisui T & Fujiwara S (1999) A retinoic acid-inducible modular protease in budding ascidians. *Dev Biol*. **214**, 38-45.
26. Yamada K, Takabatake T & Takeshima K (2000) Isolation and characterization of three novel serine protease genes from *Xenopus laevis*. *Gene*. **252**, 209-216.
27. Gough PJ, Greaves DR & Gordon S (1998) A naturally occurring isoform of the human macrophage scavenger receptor (SR-A) gene generated by alternative splicing blocks modified LDL uptake. *J Lipid Res*. **39**, 531-543.
28. Sankala M, Brannstrom A, Schulthess T, Bergmann U, Morgunova E, Engel J, Tryggvason K & Pikkarainen T (2002) Characterization of recombinant soluble macrophage scavenger receptor MARCO. *J Biol Chem*. **277**, 33378-33385.
29. Jourdan-Le Saux C, Tronecker H, Bogic L, Bryant-Greenwood GD, Boyd CD & Csiszar K (1999) The LOXL2 gene encodes a new lysyl oxidase-like protein and is expressed at high levels in reproductive tissues. *J Biol Chem*. **274**, 12939-12944.
30. Jourdan-Le Saux C, Tomsche A, Ujfalusi A, Jia L & Csiszar K (2001) Central nervous system, uterus, heart, and leukocyte expression of the LOXL3 gene, encoding a novel lysyl oxidase-like protein. *Genomics*. **74**, 211-218.
31. Asuncion L, Fogelgren B, Fong KS, Fong SF, Kim Y & Csiszar K (2001) A novel human lysyl oxidase-like gene (LOXL4) on chromosome 10q24 has an altered scavenger receptor cysteine rich domain. *Matrix Biol*. **20**, 487-491.
32. Proba K, Gschwend TP & Sonderegger P (1998) Cloning and sequencing of the cDNA encoding human neurotrypsin. *BBA-Gene Struct Expr*. **1396**, 143-147.
33. Wolfer DP, Lang R, Cinelli P, Madani R & Sonderegger P (2001) Multiple roles of neurotrypsin in tissue morphogenesis and nervous system development suggested by the mRNA expression pattern. *Mol Cell Neurosci*. **18**, 407-433.
34. Klezovitch O, Chevillet J, Mirosevich J, Roberts RL, Matusik RJ & Vasioukhin V (2004) Hepsin promotes prostate cancer progression and metastasis. *Cancer Cell*. **6**, 185-195.
35. Kitamoto Y, Yuan X, Wu Q, McCourt DW & Sadler JE (1994) Enterokinase, the initiator of intestinal digestion, is a mosaic protease composed of a distinctive assortment of domains. *Proc Natl Acad Sci USA*. **91**, 7588-7592.

36. Paoloni-Giacobino A, Chen H, Peitsch MC, Rossier C & Antonarakis SE (1997) Cloning of the TMPRSS2 gene, which encodes a novel serine protease with transmembrane, LDLRA, and SRCR domains and maps to 21q22.3. *Genomics*. **44**, 309-320.

37. Vaarala MH, Porvari KS, Kellokumpu S, Kyllonen AP & Vihko PT (2001) Expression of transmembrane serine protease TMPRSS2 in mouse and human tissues. *J Pathol.* **193**, 134-140.

38. Guipponi M, Vuagniaux G, Wattenhofer M, Shibuya K, Vazquez M, Dougherty L, Scamuffa N, Guida E, Okui M, Rossier C, Hancock M, Buchet K, Reymond A, Hummler E, Marzella PL, Kudoh J, Shimizu N, Scott HS, Antonarakis SE & Rossier BC (2002) The transmembrane serine protease (TMPRSS3) mutated in deafness DFNB8/10 activates the epithelial sodium channel (ENaC) in vitro. *Hum Mol Genet.* **11**, 2829-2836.

39. Wallrapp C, Hahnel S, Muller-Pillasch F, Burghardt B, Iwamura T, Ruthenburger M, Lerch MM, Adler G & Gress TM (2000) A novel transmembrane serine protease (TMPRSS3) overexpressed in pancreatic cancer. *Cancer Res.* **60**, 2602-2606.

40. Yamaguchi N, Okui A, Yamada T, Nakazato H & Mitsui S (2002) Spinesin/TMPRSS5, a novel transmembrane serine protease, cloned from human spinal cord. *J Biol Chem.* **277** , 6806-6812.

41. Yan W, Wu F, Morser J & Wu Q (2000) Corin, a transmembrane cardiac serine protease, acts as a pro-atrial natriuretic peptide-converting enzyme. *Proc Natl Acad Sci USA.* **97**, 8525-8529.

42. Ullman CG, Chamberlain D, Ansari A, Emery VC, Haris PI, Sim RB & Perkins SJ (1998) Human complement factor I: its expression by insect cells and its biochemical and structural characterisation. *Mol Immunol.* **35**, 503-512.

43. Hellstern S, Sasaki T, Fauser C, Lustig A, Timpl R & Engel J (2002) Functional studies on recombinant domains of Mac-2-binding protein. *J Biol Chem.* **277**, 15690-15696.

44. Jalkanen K, Leu T, Bono P, Salmi M, Jalkanen S & Smith DJ (2001) Distinct ligand binding properties of Mac-2-binding protein and mousephilin C-associated protein. *Eur J Immunol.* **31**, 3075-3084.

45. Friedman J, Trahey M & Weissman I (1993) Cloning and characterization of cyclophilin C-associated protein: a candidate natural cellular ligand for cyclophilin C. *Proc Natl Acad Sci USA.* **90**, 6815-6819.

46. Trahey M & Weissman IL (1999) Cyclophilin C-associated protein: A normal secreted glycoprotein that down-modulates endotoxin and proinflammatory responses in vivo. *Proc Natl Acad Sci USA.* **96**, 3006-3011.

47. Chicheportiche Y & Vassalli P (1994) Cloning and expression of a mouse macrophage cDNA coding for a membrane glycoprotein of the scavenger receptor cysteine-rich domain family. *J Biol Chem.* **269**, 5512-5517.

48. Pahler S, Blumbach B, Muller I & Muller WE (1998) Putative multiadhesive protein from the marine sponge *Geodia cydonium*: cloning of the cDNA encoding a fibronectin-, an SRCR-, and a complement control protein module. *J Exp Zool.* **282**, 332-343.

49. Mayer WE & Tichy H (1995) A cDNA clone from the sea lamprey *Petromyzon marinus* coding for a scavenger receptor Cys-rich (SRCR) domain protein. *Gene.* **164**, 267-271.

50. Iwasaki K, Morimatsu M, Inanami O, Uchida E, Syuto B, Kuwabara M & Niiyama M (2001) Isolation, characterization, and cDNA cloning of chicken turpentine-induced protein, a new member of the scavenger receptor cysteine-rich (SRCR) family of proteins. *J Biol Chem.* **276**, 9400-9405.

51. Gebe JA, Llewellyn M, Hoggatt H & Aruffo A (2000) Molecular cloning, genomic organization and cell-binding characteristics of mouse Spalpha. *Immunology*. **99**, 78-86.
52. Padilla O, Pujana MA, Lopez-de la Iglesia A, Gimferrer I, Arman M, Vila JM, Places L, Vives J, Estivill X & Lozano F (2002) Cloning of S4D-SRCRB, a new soluble member of the group B scavenger receptor cysteine-rich family (SRCR-SF) mapping to human chromosome 7q11.23. *Immunogenetics*. **54**, 621-634.
53. Gronlund J, Vitved L, Lausen M, Skjodt K & Holmskov U (2000) Cloning of a navel scavenger receptor cysteine-rich type I transmembrane molecule (M160) expressed by human macrophages. *J Immunol*. **165**, 6406-6415.
54. Hanby-Flarida MD, Trask OJ, Yang TJ & Baldwin CL (1996) Modulation of WC1, a lineage-specific cell surface molecule of gamma/delta T cells augments cellular proliferation. *Immunology*. **88**, 116-123.
55. Kirkham PA, Takamatsu HH, Lam EW & Parkhouse RM (2000) Ligation of the WC1 receptor induces gamma delta T cell growth arrest through fumonisin B1-sensitive increases in cellular ceramide. *J Immunol*. **165**, 3564-3570.
56. Kang W & Reid KB (2003) DMBT1, a regulator of mucosal homeostasis through the linking of mucosal defense and regeneration? *FEBS Lett*. **540**, 21-25.
57. Thornton DJ, Davies JR, Kirkham S, Gautrey A, Khan N, Richardson PS & Sheehan JK (2001) Identification of a nonmucin glycoprotein (gp-340) from a purified respiratory mucin preparation: evidence for an association involving the MUC5B mucin. *Glycobiology*. **11**, 969-977.
58. Ligtenberg AJ, Bikker FJ, Blieck-Hogervorst JM, Veerman EC & Nieuw Amerongen AV (2004) Binding of Salivary Agglutinin (SAG) to IgA. *Biochem J*. **Pt.**.
59. Cheng H, Bjerknes M & Chen H (1996) CRP-ductin: a gene expressed in intestinal crypts and in pancreatic and hepatic ducts. *Anat Rec*. **244**, 327-343.
60. Thim L & Mortz E (2000) Isolation and characterization of putative trefoil peptide receptors. *Regul Pept*. **90**, 61-68.
61. Madsen J, Tornoe I, Nielsen O, Lausen M, Krebs I, Mollenhauer J, Kollender G, Poustka A, Skjodt K & Holmskov U (2003) CRP-ductin, the mouse homologue of gp-340/deleted in malignant brain tumors 1 (DMBT1), binds gram-positive and gram-negative bacteria and interacts with lung surfactant protein D. *Eur J Immunol*. **33**, 2327-2336.
62. De Lisle RC & Ziemer D (2000) Processing of pro-Muclin and divergent trafficking of its products to zymogen granules and the apical plasma membrane of pancreatic acinar cells. *Eur J Cell Biol*. **79**, 892-904.
63. Matsushita F, Miyawaki A & Mikoshiba K (2000) Vomeroglandin/CRP-ductin is strongly expressed in the glands associated with the mouse vomeronasal organ: Identification and characterization of mouse vomeroglandin. *Biochem Biophys Res Co*. **268**, 275-281.
64. Bisgaard HC, Holmskov U, Santoni-Rugiu E, Nagy P, Nielsen O, Ott P, Hage E, Dalhoff K, Rasmussen LJ & Tygstrup N (2002) Heterogeneity of ductular reactions in adult rat and human liver revealed by novel expression of deleted in malignant brain tumor 1. *Am J Pathol*. **161**, 1187-1198.
65. Li XJ & Snyder SH (1995) Molecular cloning of Ebnerin, a von Ebner's gland protein associated with taste buds. *J Biol Chem*. **270**, 17674-17679.

66. Takito J, Hikita C & Al Awqati Q (1996) Hensin, a new collecting duct protein involved in the in vitro plasticity of intercalated cell polarity. *J Clin Invest.* **98**, 2324-2331.
67. Al Awqati Q, Vijayakumar S & Takito J (2003) Terminal differentiation of epithelia from trophectoderm to the intercalated cell: The role of hensin. *J Am Soc Nephrol.* **14**, S16-S21.
68. Offner GD, Nunes DP, Keates AC, Afdhal NH & Troxler RF (1998) The amino-terminal sequence of MUC5B contains conserved multifunctional D domains: implications for tissue-specific mucin functions. *Biochem Biophys Res Commun.* **251**, 350-355.
69. Tordai H, Banyai L & Patthy L (1999) The PAN module: the N-terminal domains of plasminogen and hepatocyte growth factor are homologous with the apple domains of the prekallikrein family and with a novel domain found in numerous nematode proteins. *FEBS Lett.* **461**, 63-67.
70. Aravind L & Koonin EV (1999) Fold prediction and evolutionary analysis of the POZ domain: structural and evolutionary relationship with the potassium channel tetramerization domain. *J Mol Biol.* **285**, 1353-1361.
71. Cambi A & Figdor CG (2003) Dual function of C-type lectin-like receptors in the immune system. *Curr Opin Cell Biol.* **15**, 539-546.
72. Tjoelker LW, Gosting L, Frey S, Hunter CL, Trong HL, Steiner B, Brammer H & Gray PW (2000) Structural and functional definition of the human chitinase chitin-binding domain. *J Biol Chem.* **275**, 514-520.
73. Chamoy L, Nicolai M, Quennedey B, Gaill F & Delachambre J (2000) Characterization of a cDNA encoding RP43, a CUB-domain-containing protein from the tube of Riftia pachyptila (Vestimentifera), and distribution of its transcript. *Biochem J.* **350 Pt 2**, 421-427.
74. Gregory LA, Thielens NM, Arlaud GJ, Fontecilla-Camps JC & Gaboriaud C (2003) X-ray structure of the Ca2+-binding interaction domain of C1s. Insights into the assembly of the C1 complex of complement. *J Biol Chem.* **278**, 32157-32164.
75. Lin HH, Stacey M, Saxby C, Knott V, Chaudhry Y, Evans D, Gordon S, McKnight AJ, Handford P & Lea S (2001) Molecular analysis of the epidermal growth factor-like short consensus repeat domain-mediated protein-protein interactions: dissection of the CD97-CD55 complex. *J Biol Chem.* **276**, 24160-24169.
76. Appella E, Weber IT & Blasi F (1988) Structure and function of epidermal growth factor-like regions in proteins. *FEBS Lett.* **231**, 1-4.
77. Chamberlain D, Ullman CG & Perkins SJ (1998) Possible arrangement of the five domains in human complement factor I as determined by a combination of X-ray and neutron scattering and homology modeling. *Biochemistry-US.* **37**, 13918-13929.
78. Akiyama SK (1996) Integrins in cell adhesion and signaling. *Hum Cell.* **9**, 181-186.
79. Dierick H & Bejsovec A (1999) Cellular mechanisms of wingless/Wnt signal transduction. *Curr Top Dev Biol.* **43**, 153-190.
80. Barclay AN (2003) Membrane proteins with immunoglobulin-like domains--a master superfamily of interaction molecules. *Semin Immunol.* **15**, 215-223.
81. Yamamura Y, Yamashiro K, Tsuruoka N, Nakazato H, Tsujimura A & Yamaguchi N (1997) Molecular cloning of a novel brain-specific serine protease with a kringle-like structure and three scavenger receptor cysteine-rich motifs. *Biochem Biophys Res Commun.* **%20;239**, 386-392.
82. Trexler M, Banyai L & Patthy L (2000) The LCCL module. *Eur J Biochem.* **267**, 5751-5757.

83. Jovine L, Qi H, Williams Z, Litscher E & Wassarman PM (2002) The ZP domain is a conserved module for polymerization of extracellular proteins. *Nat Cell Biol.* **4**, 457-461.

84. Doi T, Higashino K, Kurihara Y, Wada Y, Miyazaki T, Nakamura H, Uesugi S, Imanishi T, Kawabe Y, Itakura H & . (1993) Charged collagen structure mediates the recognition of negatively charged macromolecules by macrophage scavenger receptors. *J Biol Chem.* **268**, 2126-2133.

85. Brannstrom A, Sankala M, Tryggvason K & Pikkarainen T (2002) Arginine residues in domain V have a central role for bacteria-binding activity of macrophage scavenger receptor MARCO. *Biochem Biophys Res Commun.* **290**, 1462-1469.

86. Bikker FJ, Ligtenberg AJ, Nazmi K, Veerman EC, van't Hof W, Bolscher JG, Poustka A, Nieuw Amerongen AV & Mollenhauer J (2002) Identification of the bacteria-binding peptide domain on salivary agglutinin (gp-340/DMBT1), a member of the scavenger receptor cysteine-rich superfamily. *J Biol Chem.* **277**, 32109-32115.

87. Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, MarkleY JL & Sykes BD (1995) $^{1}H$, $^{13}C$ and $^{15}N$ chemical-shift referencing in biomolecular NMR. *J Biomol NMR.* **6**, 135-140.

88. Yamazaki T, Lee W, Arrowsmith CH, Muhandiram DR & Kay LE (1994) A suite of triple-resonance NMR experiments for the backbone assignment of $^{15}N$, $^{13}C$, $^{2}H$ labeled proteins with high- sensitivity. *J Am Chem Soc.* **116**, 11655-11666.

89. Wittekind M & Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha-carbon and beta-carbon resonances in proteins. *J Magn Reson Ser B.* **101**, 201-205.

90. Muhandiram DR & Kay LE (1994) Gradient-enhanced triple-resonance 3-dimensional NMR experiments with improved sensitivity. *J Magn Reson Ser B.* **103**, 203-216.

91. Boucher W, Laue ED, Campbellburk S & Domaille PJ (1992) 4-dimensional heteronuclear triple resonance NMR methods for the assignment of backbone nuclei in proteins. *J Am Chem Soc.* **114**, 2262-2264.

92. Marion D, Ikura M, Tschudin R & Bax A (1989) Rapid recording of 2D NMR-spectra without phase cycling - application to the study of hydrogen-exchange in proteins. *J Magn Res.* **85**, 393-399.

93. Bax A, Clore GM & Gronenborn AM (1990) $^{1}H$-$^{1}H$ correlation via isotropic mixing of $^{13}C$ magnetization, a new 3-dimensional approach for assigning $^{1}H$ and $^{13}C$ spectra of $^{13}C$ enriched proteins. *J Magn Res.* **88**, 425-431.

94. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J & Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* **6**, 277-293.

95. Boucher, W (2002) AZARA suite for NMR data (version 2.7) Department of Biochemistry, University of Cambridge.

96. Kraulis PJ (1989) ANSIG - A program for the assignment of protein H-1 2D-NMR spectra by interactive computer-graphics. *Journal of Magnetic Resonance.* **84**, 627-633.

97. Cornilescu G, Delaglio F & Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J Biomol NMR.* **13**, 289-302.

98. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T & Warren GL (1998)

Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D.* **54**, 905-921.

99. Linge JP & Nilges M (1999) Influence of non-bonded parameters on the quality of NMR structures: A new force field for NMR structure calculation. *J Biomol NMR.* **13**, 51-59.
100. Linge JP, Williams MA, Spronk CAEM, Bonvin AMJJ & Nilges M (2003) Refinement of protein structures in explicit solvent. *Proteins.* **50**, 496-506.
101. Folmer RHA, Konings RNH, Hilbers CW & Nilges M (1997) Floating stereospecific assignment revisited: Application to a 18 kDa protein and comparison with J-coupling data. *J Biomol NMR.* **9**, 245-258.
102. Wakefield RID, Smith BO, Nan XS, Free A, Soteriou A, Uhrin D, Bird AP & Barlow PN (1999) The solution structure of the domain from MeCP2 that binds to methylated DNA. *J Mol Biol.* **291**, 1055-1065.
103. Levitt MH (2001) Spin dynamics: Basics of nuclear magnetic resonance. John Wiley & Sons, Ltd., Chichester, West Sussex.
104. Case DA (1995) Calibration of ring-current effects in proteins and nucleic acids. *J Biomol NMR.* **6**, 341-346.
105. Koradi R, Billeter M & Wuthrich K (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph.* **14**, 51-55.
106. Ramboarina S & Redfield C (2003) Structural characterisation of the human alpha-lactalbumin molten globule at high temperature. *J Mol Biol.* **330**, 1177-1188.
107. Kim S, Bracken C & Baum J (1999) Characterization of millisecond time-scale dynamics in the molten globule state of alpha-lactalbumin by NMR. *J Mol Biol.* **294**, 551-560.
108. Yao J, Dyson HJ & Wright PE (1997) Chemical shift dispersion and secondary structure prediction in unfolded and partly folded proteins. *FEBS Lett.* **419**, 285-289.
109. Sharon N & Lis H (2001) The structural basis for carbohydrate recognition by lectins. *Adv Exp Med Biol.* **491:1-16.**, 1-16.
110. Helenius A & Aebi M (2001) Intracellular functions of N-linked glycans. *Science.* **291**, 2364-2369.
111. Rudd PM, Wormald MR, Stanfield RL, Huang M, Mattsson N, Speir JA, DiGennaro JA, Fetrow JS, Dwek RA & Wilson IA (1999) Roles for glycosylation of cell surface receptors involved in cellular immune recognition. *J Mol Biol.* **293**, 351-366.
112. Wormald MR & Dwek RA (1999) Glycoproteins: glycan presentation and protein-fold stability. *Struct Fold Des.* **7**, R155-R160.
113. Grey MJ, Wang C & Palmer AG, III (2003) Disulfide bond isomerization in basic pancreatic trypsin inhibitor: multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J Am Chem Soc.* **125**, 14324-14335.
114. Beeser SA, Oas TG & Goldenberg DP (1998) Determinants of backbone dynamics in native BPTI: cooperative influence of the 14-38 disulfide and the Tyr35 side-chain. *J Mol Biol.* **284**, 1581-1596.
115. Otting G, Liepinsh E & Wuthrich K (1993) Disulfide bond isomerization in BPTI and BPTI(G36S): an NMR study of correlated mobility in proteins. *Biochemistry-US.* **32**, 3571-3582.
116. Szyperski T, Luginbuhl P, Otting G, Guntert P & Wuthrich K (1993) Protein dynamics studied by rotating frame 15N spin relaxation times. *J Biomol NMR.* **3**, 151-164.

117. Sorensen MD, Bjorn S, Norris K, Olsen O, Petersen L, James TL & Led JJ (1997) Solution structure and backbone dynamics of the human alpha3-chain type VI collagen C-terminal Kunitz domain. *Biochemistry-US.* **36**, 10439-10450.

118. Li YC & Montelione GT (1995) Human type-alpha transforming growth factor undergoes slow conformational exchange between multiple backbone conformations as characterized by nitrogen-15 relaxation measurements. *Biochemistry-US.* **34**, 2408-2423.

119. Fairbrother WJ, Liu J, Pisacane PI, Sliwkowski MX & Palmer AG, III (1998) Backbone dynamics of the EGF-like domain of heregulin-alpha. *J Mol Biol.* **279**, 1149-1161.

120. Bretthauer RK & Castellino FJ (1999) Glycosylation of Pichia pastoris-derived proteins. *Biotechnol Appl Biochem.* **30 ( Pt 3)**, 193-200.

121. Rudd PM, Wormald MR, Harvey DJ, Devasahayam M, McAlister MS, Brown MH, Davis SJ, Barclay AN & Dwek RA (1999) Oligosaccharide analysis and molecular modeling of soluble forms of glycoproteins belonging to the Ly-6, scavenger receptor, and immunoglobulin superfamilies expressed in Chinese hamster ovary cells. *Glycobiology.* **9**, 443-458.

122. Hohenester E, Sasaki T & Timpl R (1999) Crystal structure of a scavenger receptor cysteine-rich domain sheds light on an ancient superfamily. *Nat Struct Biol.* **6**, 228-232.

123. Somoza JR, Ho JD, Luong C, Ghate M, Sprengeler PA, Mortara K, Shrader WD, Sperandio D, Chan H, McGrath ME & Katz BA (2003) The structure of the extracellular region of human hepsin reveals a serine protease domain and a novel scavenger receptor cysteine-rich (SRCR) domain. *Structure.* **11**, 1123-1131.

124. Kabsch W & Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* **22**, 2577-2637.

125. Skonier JE, Bodian DL, Emswiler J, Bowen MA, Aruffo A & Bajorath J (1997) Mutational analysis of the CD6 ligand binding domain. *Protein Eng.* **10**, 943-947.

126. Bodian DL, Skonier JE, Bowen MA, Neubauer M, Siadak AW, Aruffo A & Bajorath J (1997) Identification of residues in CD6 which are critical for ligand binding. *Biochemistry.* **36**, 2637-2641.

127. Bajorath J (1999) Three-dimensional analysis of CD6 site-directed mutagenesis and monoclonal antibody binding studies using the x-ray structure of mac-2 binding protein and a molecular model of the CD6 ligand binding domain. *J Mol Model.* **5**, 263-270.

128. Bowen MA, Aruffo AA & Bajorath J (2000) Cell surface receptors and their ligands: in vitro analysis of CD6-CD166 interactions. *Proteins.* **40**, 420-428.

# Chapter Five

## General conclusions and perspectives

*A systematic screen for suitable folding conditions makes in vitro protein folding a highly effective tool for successful recombinant protein production*

Recombinant proteins that express in inclusion bodies are a common occurrence in structural biology projects. In large scale projects, around 30-40% of the targets that express well in *Escherichia coli*. accumulate in inclusion bodies; a similar percentage has been reported for single-domain cytoplasmic yeast proteins[1], human protein domains[2] and non-membrane proteins from thermophilic archaebacteria[3].

Considerable effort has been expended in trying to reduce or avoid inclusion body formation and produce protein in a soluble form. A variety of strategies have been developed including: directed evolution to create soluble mutants[4], direction of the recombinant protein to the periplasmic space of *E. coli*, co-expression with chaperones[5] and use of cell-free expression systems[6]. A recent publication has compared the utility of three different methods, in respect to their ability to produce soluble protein, using 20 different proteins that expressed in inclusion bodies under standard conditions[1]. The authors concluded that none of the studied methods (chaperone co-expression, cell-free expression and protein *in vitro* folding) had a superior performance over the others and that the best method for a given protein cannot be predicted. These different methods should be regarded as complementary and not competing techniques.

*In vitro* protein folding is the only strategy for obtaining folded protein that takes advantage of the benefits of inclusion body expression instead of trying to avoid their formation. These advantages include: high expression yield, reduced toxicity for the expression host and reduced degradation during expression and purification. *In vitro* folding is also well suited for the preparation of proteins that require co-factors, prosthetic groups, metals or protein partners to attain their 3D structure, and it is compatible with the use of protein chaperones in order to further improve yields.

The major disadvantage of *in vitro* protein folding is the difficulty of finding suitable folding conditions given that optimal conditions vary from one protein to another. However, the development of efficient methods to identify the optimal folding conditions for any particular protein will make *in vitro* folding amenable to implementation on projects of all scales. The fractional factorial screen for *in vitro* protein folding presented in this project, and similar screens published previously, provide strong evidence of the merits of this technique in finding optimal folding conditions. Fractional factorial screens are systematic, easy to implement, readily modifiable to suit particular circumstances, fast, and suitable for small and large scale projects.

The *in vitro* oxidative protein folding screen developed in this project succeeded in identifying suitable conditions to fold RNAse A, lysozyme, and two of the SRCR domains of CD5. However, the general efficacy of the screen needs to be further demonstrated by assessing a larger number of proteins. In an immediate future, the obvious candidates within our laboratory are the remaining SRCR domain of CD5 (CD5d2) and a protein construct comprising both CD5d2 and CD5d3 (CD5d2+3). I have subcloned CD5d2+3 and CD5d2 into bacterial expression vectors. While CD5d2+3 expresses in good yield in inclusion bodies, the current CD5d2 construct does not show any expression. New constructs of CD5d2 need to be made in order to overcome this problem, and it is possible that expression can be attained by changing the expression vector or by modifying the length of the construct. Our initial interest in creating CD5d2+d3 arose because CD5d2 and CD5d3 do not have a clear linker region between them (such as the stalk region between CD5d1 and CD5d2) and it seemed possible that they might require each other to attain their native fold. At this time, the $^{1}H$-$^{15}N$ HSQC spectrum of CD5d3 (Figure II.13) strongly suggests that CD5d3 does not require CD5d2 for attainment of its 3D structure. However, tandem SRCR domains are a very common feature of the members of the superfamiliy (Figures IV.1-3) and a CD5d2+3 structure

would be the first example of an arrangement of tandem SRCR domains, and would perhaps provide information about how SRCR domains interact with each other.

Another imminent target with which to test the utility of the *in vitro* folding screen is the Ly-6 domain of the GPI anchored protein Prod1. Structural studies of Prod1 are an ongoing collaboration with Prof Jeremy Brockes at the Department of Biochemistry of UCL. Prod1 is a single-domain glycosylated cell surface protein implicated in cellular interactions during newt limb regeneration. Ly-6 domains are mainly beta folds of *ca.* 100 residue-long and contain five conserved disulphide bonds. Efforts are being made in order to express the folded domain *in vivo* using the thioredoxin reductase and glutathione reductase suppressor strain Origami, which it is known to enhance disulphide bond formation in the cytoplasm[7]. However, although some soluble protein does accumulate, most of the expressed protein is in inclusion bodies, making the Ly-6 domain of Prod1 another case for the application of the *in vitro* folding screen.

The oxidative folding screen presented here would benefit from further improvements in the methodology, particularly in order to reduce sample handling. A disadvantage of the current implementation is that it includes a dialysis step that requires extensive sample manipulation and relatively large volumes (of the order of a few millilitres). This dialysis step is performed in order to remove guanidine HCl and arginine, which might otherwise enhance the solubility of protein aggregates. However, it would be worth investigating whether this step is essential for successful determination of suitable folding conditions. It may be that determining folding yield in the presence of arginine or guanidine HCl gives similar results and the dialysis is only required for preparative purposes.

The implementation of hydrophobic interaction chromatography (HIC), instead of reversed-phase chromatography, for quantifying folding yields may have several advantages. Folding yield quantification by HIC would make the screen suitable for non disulphide-bonded proteins and would avoid the generation of organic solvent waste. Another approach that requires consideration is whether a measurement of protein solubility, instead of folding yield, would be sufficient to determine optimal folding conditions, as this would save a considerate amount of time and effort. Turbidity measurements (determination of absorbance at 390 or 340 nm) have recently been found to be a suitable method to evaluate *in vitro* folding conditions in high throughput screens[1,8].

*Orthologue sequence comparison is an efficient and straightforward strategy for the design of mutations in order to improve solubility and stability*

Low solubility of proteins is a common problem in structural biology, as studies often require protein concentrations of the order of 1 mM. Approaches that have been used to generate versions of proteins with higher solubility include directed evolution[4] and the use of solubility enhancing tags[9]. However, the approach utilised here was to make rational replacements of a limited number of apparently solvent exposed residues, replacing apolar with polar residues. Since a folded protein is always in equilibrium with its unfolded or partially unfolded forms, improved behaviour of the protein as a result of mutation may arise either from a reduction in aggregation of the folded protein or its (partially) unfolded forms, or via an increase in the stability of the soluble folded form (which would diminish the proportion of any aggregation prone unfolded form present in solution). Given this variety of ways in which mutation may alter solubility, it is not always an easy task to locate good candidate residues for mutation.

Enhancement of the stability of the folded form of the protein is one route to improving yield of soluble protein. It is usually the case that the largest contribution to stability arises from burial of hydrophobic residues in the protein core. Unfortunately, it is difficult to make stability improving mutations in the well-packed and probably evolutionarily optimised protein interior. However, it has been observed that the surface of thermophilic proteins comprises more charged amino acid residues than in their mesophilic orthologues[10] and has a larger fractional polar content[11]. The increased stability of many thermophilic proteins has been found to be a consequence of a small number of differences in the solvent-exposed residues between the thermophile and its mesophilic orthologues[12]. It is hypothesised that short and long-range electrostatic interactions between solvent-exposed charged amino acids residues make an important contribution to this stability[13-17] by destabilising the unfolded state and by stabilising the folded state[18] [12,19]. This suggests that apolar-to-polar and especially apolar-to-charged amino acid mutations of solvent-exposed residues are likely to be successful in improving solubility (and protein yield) by improving stability.

In a recent high throughput study on prokaryote proteins, it was estimated that the probability of obtaining one sample suitable for structural studies using three orthologue sequences instead of the sequence from a single organism rises from 70% to 96%[20]. Mutation of one or two exposed amino acid residues has been proved to be sufficient to significantly enhance the stability of several proteins[21]. Thus, exploiting the subtle sequence variation among orthologues in selecting mutations is sufficient to alter the stability and solubility of proteins. In this project, I have provided new evidence to support the use of a multiple sequence alignment with orthologue proteins as basis for selecting mutations that are likely to improve the protein's behaviour. In the case of *h*CD5d1, this strategy was successful in engineering proteins with superior properties for structural studies in four out of eight attempts. *A posteriori* analysis of these mutations in light of the solved structure showed that the successful mutations V97K, L26K, V88D, and M124Q all correspond to solvent-exposed residues that do not seem to have a critical structural role. On the other hand, residues W28 and G129, whose mutants did not yield folded protein, are buried residues whose alteration probably disturbs the three dimensional structure. More intriguing are the mutations V88N and M124N, which despite being apolar to polar substitutions of solvent exposed residues gave a lower yield of soluble folded protein than the wild type. It is tempting to conclude that the difference in behaviour between the V88D and V88N mutants is due to the negative charge in Asp, which makes V88D more hydrophilic than V88N, and able to establish the charge-charge interactions common in stabilising mutations.

It has been proposed that relatively simple theoretical calculations can be used to make a rational design of mutations to increase stability[12], and the implementation of one of such methodologies has proved successful[21]. Adopting such calculations as a criterion for selecting mutations may help to improve folding yields or make proteins more amenable for structural studies. It would also provide evidence to address the question of whether there is a correlation between improvement of folding yields and enhancement of stability of proteins.

*The N-terminal SRCR domain of CD5 is an interesting target for dynamic studies using NMR*

NMR is a powerful technique to characterise molecular dynamics at atomic resolution. Using NMR it is possible not only to identify the regions of the structure that are exchanging between different conformations, but to measure the rate constant of the exchange process and the relative populations of the conformations. Techniques to characterise chemical exchange processes in the microsecond to millisecond timescale, namely the experiments based on the Carr-Purcell-Meibbom-Gill (CPMG) and the $R_{1\rho}$ relaxation dispersion methods[22,23], are not as well-established as are the experimental methods for the investigation of faster motions. The development of methods to study microsecond to millisecond timescale dynamics remains an area of active research[22-24] and the complex nature of the dynamics of *h*CD5d1 may provide an opportunity to establish and experiment with these state-of-the-art methodologies within the group.

An in-depth analysis of the residues undergoing conformational exchange and the estimation of the timescales for their motions would provide evidence to determine the nature of the mechanism of improvement of the quality of the NMR spectra with increasing temperature, *i.e.* whether it is a consequence of an increase in the rate of the exchange among different conformations or of the stabilisation of a single conformer. Furthermore, it may be possible to use knowledge of the chemical shift changes with temperature, in conjunction with such exchange analysis, to provide additional information about the structural origin of the mobility. Chemical shift changes have been previously correlated with structural changes (such as disulphide bond rotation) using data from chemical shift databases[25,26], or by measuring the relaxation of more than one nucleus[27].

The fact that *h*CD5d1 expresses in eukaryotic systems with very poor yields, mean that it is unlikely that we will be able to further compare the dynamics of the mammalian, yeast and bacterial versions in order to determine the role of different glycans in restricting flexibility. However, a technique to produce O-glycosylated proteins in *E. coli* has been recently reported[28,29], and perhaps the bacterial expression of N-glycosylated proteins will become a reality in the near future.

*The C-terminal SRCR domain of CD5 is a promising target for ligand binding experiments*

The peaks in the $^{1}H$-$^{15}N$ HSQC of the C-terminal SRCR domain of CD5 (CD5d3), are well dispersed and have a uniform intensity, as was shown in Chapter II. This result suggests that the CD5d3 polypeptide is folded and is not experiencing significant conformational exchange. Consequently, its structure should be determined in the near future and it is anticipated that it will be of higher 'quality' than that of CD5d1. Furthermore, CD5d3 does not have any putative N-glycosylation sites, which means that study of CD5d3 will not be complicated by issues with respect to the functional relevance of the structure arising from the presence/absence of glycan.

The extracellular region of CD5 has been proposed to bind the framework sequences or the variable immunoglobulin domains of many antibodies[30]. In this case, the interaction was mapped to the intermediate SRCR domain (CD5d2)[31]. The only well characterised protein-protein interaction of an SRCR domain involves the C-terminal SRCR domain of CD6 (CD6d3) and the N-terminal immunoglobulin domain of CD166[32]. As CD5 and CD6 are closely related and have a very similar domain organisation, it is feasible that CD5d3 also binds an immunoglobulin domain Studying the putative antibody and immunoglobulin binding properties of CD5 using different constructs comprising one or two of the SRCR domains of CD5, would be possible with the recombinant protein produced using methods created in this study. This could be achieved by immobilising antibodies to beads and attempting binding of the CD5 constructs to them. The opposite strategy, immobilising the CD5 constructs to the beads, might prove valuable in a general search for CD5 ligands. The CD5 construct-coated beads could be used to attempt the isolation of molecules from, for example, lymphocyte cell lysates.

Also, having implemented a strategy to produce SRCR domains of CD5 suitable for structural characterisation using a bacterial expression system, the recombinant expression of the SRCR domains of CD6 is not likely to present a major challenge. Availability of $^{15}N$-labelled CD6d3 would open the doors for the characterisation of the known CD6-CD166 interaction using NMR.

*And finally*

The increased availability of automated techniques and apparatus and the success of high-throughput methodologies are fundamentally changing the way of conducting research. Nowadays, much research must be done rapidly and cost-effectively to generate results in the short-term. What recalcitrant proteins like the SRCR domains of CD5 have to offer in this environment, is a challenge to current methods and a motivation to develop new strategies for future wider application. The lessons that have been learnt with CD5 will certainly be useful in rapidly and efficiently overcoming similar problems in future projects.

## References


1. Tresaugues L, Collinet B, Minard P, Henckes G, Aufrere R, Blondeau K, Liger D, Zhou CZ, Janin J, van Tilbeurgh H & Quevillon-Cheruel S (2004) Refolding strategies from inclusion bodies in a structural genomics project. *J Struct Funct. Genomics.* **5** , 195-204.
2. Folkers GE, van Buuren BN & Kaptein R (2004) Expression screening, protein purification and NMR analysis of human protein domains for structural genomics. *J Struct Funct Genomics.* **5**, 119-131.
3. Christendat D, Yee A, Dharamsi A, Kluger Y, Savchenko A, Cort JR, Booth V, Mackereth CD, Saridakis V, Ekiel I, Kozlov G, Maxwell KL, Wu N, McIntosh LP, Gehring K, Kennedy MA, Davidson AR, Pai EF, Gerstein M, Edwards AM & Arrowsmith CH (2000) Structural proteomics of an archaeon. *Nat Struct Biol.* **7**, 903-909.
4. Pedelacq JD, Piltch E, Liong EC, Berendzen J, Kim CY, Rho BS, Park MS, Terwilliger TC & Waldo GS (2002) Engineering soluble proteins for structural genomics. *Nat Biotechnol.* **20**, 927-932.
5. Nishihara K, Kanemori M, Kitagawa M, Yanagi H & Yura T (1998) Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in Escherichia coli. *Appl Environ Microbiol.* **64**, 1694-1699.
6. Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A & Yokoyama S (2004) Preparation of Escherichia coli cell extract for highly productive cell-free protein expression. *J Struct Funct Genomics.* **5**, 63-68.
7. Bessette PH, Aslund F, Beckwith J & Georgiou G (1999) Efficient folding of proteins with multiple disulfide bonds in the Escherichia coli cytoplasm. *Proc Natl Acad Sci USA.* **96**, 13703-13708.
8. Vincentelli R, Bignon C, Gruez A, Canaan S, Sulzenbacher G, Tegoni M, Campanacci V & Cambillau C (2003) Medium-scale structural genomics: strategies for protein expression and crystallization. *Acc Chem Res.* **36**, 165-172.
9. Zhou P, Lugovskoy AA & Wagner G (2001) A solubility-enhancement tag (SET) for NMR studies of poorly behaving proteins. *J Biomol NMR.* **20**, 11-14.

10. Fukuchi S & Nishikawa K (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. *J Mol Biol.* **309**, 835-843.
11. Vogt G, Woell S & Argos P (1997) Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol.* **269**, 631-643.
12. Sanchez-Ruiz JM & Makhatadze GI (2001) To charge or not to charge? *Trends Biotechnol.* **19**, 132-135.
13. Perl D, Mueller U, Heinemann U & Schmid FX (2000) Two exposed amino acid residues confer thermostability on a cold shock protein. *Nat Struct Biol.* **7**, 380-383.
14. Makhatadze GI, Loladze VV, Gribenko AV & Lopez MM (2004) Mechanism of thermostabilization in a designed cold shock protein with optimized surface electrostatic interactions. *J Mol Biol.* **336**, 929-942.
15. Grimsley GR, Shaw KL, Fee LR, Alston RW, Huyghues-Despointes BM, Thurlkill RL, Scholtz JM & Pace CN (1999) Increasing protein stability by altering long-range coulombic interactions. *Protein Sci.* **8**, 1843-1849.
16. Loladze VV, Ibarra-Molero B, Sanchez-Ruiz JM & Makhatadze GI (1999) Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface. *Biochemistry.* **38**, 16419-16423.
17. Tomazic SJ & Klibanov AM (1988) Why is one Bacillus alpha-amylase more resistant against irreversible thermoinactivation than another? *J Biol Chem.* **263**, 3092-3096.
18. Zhou HX & Dong F (2003) Electrostatic contributions to the stability of a thermophilic cold shock protein. *Biophys J.* **84**, 2216-2222.
19. Pace CN, Alston RW & Shaw KL (2000) Charge-charge interactions influence the denatured state ensemble and contribute to protein stability. *Protein Sci.* **9**, 1395-1398.
20. Savchenko A, Yee A, Khachatryan A, Skarina T, Evdokimova E, Pavlova M, Semesi A, Northey J, Beasley S, Lan N, Das R, Gerstein M, Arrowsmith CH & Edwards AM (2003) Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches. *Proteins.* **50**, 392-399.
21. Torrez M, Schultehenrich M & Livesay DR (2003) Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces. *Biophys J.* **85**, 2845-2853.
22. Akke M (2002) NMR methods for characterizing microsecond to millisecond dynamics in recognition and catalysis. *Curr Opin Struct Biol.* **12**, 642-647.
23. Palmer AG, III, Kroenke CD & Loria JP (2001) Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules. *Methods Enzymol.* **339**, 204-238.
24. Grey MJ, Wang C & Palmer AG, III (2003) Disulfide bond isomerization in basic pancreatic trypsin inhibitor: multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J Am Chem Soc.* **125**, 14324-14335.
25. Grey MJ, Wang C & Palmer AG, III (2003) Disulfide bond isomerization in basic pancreatic trypsin inhibitor: multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling. *J Am Chem Soc.* **125**, 14324-14335.
26. Zhang H, Neal S & Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J Biomol NMR.* **25**, 173-195.

27. Ishima R, Freedberg DI, Wang YX, Louis JM & Torchia DA (1999) Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure Fold Des.* **7**, 1047-1055.
28. Zhang Z, Gildersleeve J, Yang YY, Xu R, Loo JA, Uryu S, Wong CH & Schultz PG (2004) A new strategy for the synthesis of glycoproteins. *Science.* **303**, 371-373.
29. Xu R, Hanson S, Zhang Z, Yang YY, Schultz PG & Wong CH (2004) Site-specific incorporation of the mucin-type N-acetylgalactosamine-α-O-threonine into protein in *Escherichia coli*. *J Am Chem Soc.* **126**, 15654-15655.
30. Pospisil R, Fitts MG & Mage RG (1996) CD5 is a potential selecting ligand for B cell surface immunoglobulin framework region sequences. *J Exp Med.* **184**, 1279-1284.
31. Pospisil R, Silverman GJ, Marti GE, Aruffo A, Bowen MA & Mage RG (2000) CD5 is A potential selecting ligand for B-cell surface immunoglobulin: a possible role in maintenance and selective expansion of normal and malignant B cells. *Leuk Lymphoma.* **36**, 353-365.
32. Whitney GS, Starling GC, Bowen MA, Modrell B, Siadak AW & Aruffo A (1995) The membrane-proximal scavenger receptor cysteine-rich domain of CD6 contains the activated leukocyte cell adhesion molecule binding site. *J Biol Chem.* **270**, 18187-18190.

# Appendix A

# Experimental protocols

## Culture media

*M9 minimal medium with supplements*

- M9 salts (10X): 60 g/l $Na_2HPO_4$, 30 g/l $KH_2PO_4$, 5 g/l NaCl, $H_2O$ to 1000 ml
- Nitrogen source (10X): 10 g/l $(NH_4)_2SO_4$ (use $^{15}N$ ammonium sulphate for $^{15}N$-labelling)
- 0.1 M $CaCl_2$
- 1 M $MgCl_2$
- 0.01 M $FeSO_4$ (make fresh)
- Filter-sterilised 20% (w/v) D-glucose (use $^{13}C$ glucose for $^{13}C$-labelling)
- Micronutrient stock (10,000X): 0.4 M $H_3BO_3$, 30 mM $CoCl_2$, 10 mM $CuSO_4$, 8 mM $MnCl_2$, 10 mM $ZnSO_4$, 3mM $(NH_4)_6Mo_7O_{24}$
- Vitamin stock (1000X): 0.4 g/l choline chloride, 0.5g/l folic acid, 0.5 g/l pantothenic acid, 0.5g/l nicotinamide, 1 g/l myo-inositol, 0.5 g/l pyridoxal HCl, 0.5 g/l thiamine HCl, 0.05 g/l riboflavin, 1 g/l biotin

Instructions are given to make 1000 ml of culture, scale as appropriate.

1. Mix 100 ml of 10X M9 salts, 100 ml of 10X nitrogen source add $H_2O$ up to 1000 ml and adjust pH to 7.0-7.4. Pour buffer into two 2-litre baffled flasks (500 ml of buffer each). Sterilise by autoclaving for 15 minutes.

2. Mix 2 ml of 1M $MgCl_2$, 1 ml of 0.1M $CaCl_2$, 1 ml of 0.01 M $FeSO_4$, 1 ml of vitamin stock, 0.1 ml of micronutrient stock and 25 ml of 20% glucose (10 ml only if using $^{13}C$-labelled glucose). Sterilise by filtration and add half of the mixture into each of the 2-litre baffled flasks with sterile buffer.

*Luria-Bertani (LB) medium*

- 10 g/l tryptone
- 5 g/l yeast extract
- 5 g/l NaCl
- Distilled $H_2O$ up to 1 litre

Adjust pH to 7.5 using 5 M NaOH and sterilise by autoclaving.

*LB Agar plates*

- 10 g/l tryptone
- 5 g/l yeast extract
- 10 g/l NaCl
- 15 g/l agar
- Distilled $H_2O$ up to 1 litre

Adjust pH to 7.5 using 5 M NaOH and sterilise by autoclaving. Allow to cool to ~50°C, add antibiotic, mix thoroughly by swirling and pour into Petri dishes.

*SOC medium* (100 ml)

- 2 g tryptone
- 0.5 g yeast extract
- 0.06 g NaCl (or 1ml of 1 M NaCl)
- 0.02 g KCl (or 250 µl of 1 M KCl)
- Distilled $H_2O$ up to 90 ml

Adjust pH to 7.5 using 5 M NaOH and sterilise by autoclaving. Allow to cool and add a filter-sterilised solution of the following:

- 1 ml of 1 M $MgCl_2$
- 1 ml of 1 M $MgSO_4$
- 2 ml of 20% (w/v) glucose

Add sterile distilled $H_2O$ up to 100 ml and make 10-ml aliquots. Store at 20°C.

## Preparation of competent *E. coli* cells

1. Streak cells from a glycerol stock on an LB agar plate and incubate overnight at 37°C.
2. Select a well-isolated colony from the plate and inoculate 100 ml of sterile LB medium in a 250-ml baffled-flask.
3. Incubate at 37°C with vigorous shaking until the absorbance at 600 nm reaches 0.4.
4. Transfer the culture into 2 sterile disposable 50-ml conical centrifuge tubes and cool on ice for 10 minutes.
5. Centrifuge at 4000 rpm at 4°C for 10 minutes.
6. Decant the medium and invert the tubes for 1 minute to remove as much medium as possible.
7. Resuspend each of the two pellets in 10 ml of sterile ice-cold 0.1 M $CaCl_2$ and incubate on ice for 10 minutes.
8. Centrifuge at 2500 rpm at 4°C for 5 minutes and decant the medium.
9. Repeat steps 7 and 8 twice (for a total of 3 incubations in $CaCl_2$).
10. Resuspend each of the two pellets in 2 ml of ice-cold sterile 0.1 M $CaCl_2$.
11. Add 0.5 ml of 50% (w/v) glycerol to each 2-ml suspension (to achieve a final concentration of ~10% glycerol).
12. Prepare 200-µl aliquots in sterile screw-cap vials and store at -80°C.

## Transformation of competent *E. coli* cells

1. Remove the vials with competent cells from the ultrafreezer and allow them to thaw by leaving them on-ice.
2. Transfer 100-µl of the competent cell suspension into a cold sterile 1.5 ml microcentrifuge tube.
3. Add 5 µl (or less) of the plasmid DNA to be transformed into the cells and mix carefully by tapping the tube.
4. Keep on ice for 30 minutes.
5. Transfer the tube to a 42°C water bath for 90 seconds.
6. Return the tube to the ice for 2 minutes.
7. Add 900 µl of LB or SOC medium and incubate the cells for 1 hour at 37°C with occasional manual shaking.
8. Plate out 50 ml of the culture on LB agar plates containing the appropriate antibiotic and incubate overnight at 37°C.

## Preparation of *E. coli* glycerol stocks

1. Select a well-isolated bacterial colony from a freshly-streaked plate to inoculate 20 ml of sterile LB medium with the appropriate antibiotic in a sterile disposable 50-ml conical centrifuge tube (Falcon tube).
2. Incubate at 37°C with vigorous shaking until the Abs 600 nm reaches 0.6 (6-8 hrs).
3. Add 0.3 ml of 50% (w/v) glycerol to a 2 ml screw-cap vial and sterilise by autoclaving.
4. Add 0.7 ml of the log-phase *E. coli* culture to the screw-cap vial and vortrex vigorously.
5. Store at -80°C.
   - Use frozen by scraping some flakes from the surface of the stock with an inoculating loop, avoid thawing.

## DNA electrophoresis in agarose gels

- Tris-acetate-EDTA (TAE) buffer (50X): 121 g of Tris base (2 M), 29 ml of acetic acid (1 M), 9.3 g $Na_2EDTA$ (0.05 M), add distilled $H_2O$ to 500 ml, (pH ~8.5).
- Sample loading solution (6X): 30% (w/v) glycerol with a tiny amount of bromophenol blue (suitable for fragments less that 1 kb long) or xylene cyanol FF (suitable for fragments of more than 4 kb).

1. Prepare the agarose slurry by mixing the appropriate amount of solid agarose and 100 ml of 1X TAE in a 250 ml autoclavable glass bottle with a screw cap. Use 1% (w/v) agarose for plasmid DNA (>4 kb) and 2% (w/v) agarose gels for PCR products (<1 kb).
2. Heat the agarose slurry in the microwave (with the cap loose) until the agarose is dissolved.
3. Cool the agarose to ~60°C and pour the solution onto the gel tray, insert the comb and let the gel to set.
4. Transfer the gel to the electrophoresis tank and fill the tank with just enough 1X TAE to cover the gel.
5. Mix DNA samples with the sample loading solution and load into the gel.
6. Run at 50-100 V until the dye reaches the bottom of the gel.
7. Stain by immersing the gel in 100 ml of a 2 μg/ml solution of ethidium bromide in water.
8. Rinse the gel with water before visualisation.

## *In vitro* amplification of DNA by polymerase chain reaction (PCR)

*Amplification of DNA fragments*

- DNA fragments for cloning should be amplified using a high-fidelity polymerase, such as *Pyrococcus furiosus* DNA polymerase Pfu (Stratagene).

1. Program the following cycling parameters in the thermal cycler:
   01. Lid on 110°C
   02. 94°C for 5 minutes
   03. Loop 30x (
   04. 95°C for 30 seconds
   05. 55°C for 45 seconds
   06. 72°C for 2 minutes
   07. ) Loop
   08. 72°C for 5 minutes
   09. Lid off
   10. Ramp 4°C 1.0 °C/s
   11. 4°C Forever

2. Prepare the reaction mixture in an ice-chilled thin-walled PCR tube, add the components in the following order:
   - Distilled $H_2O$ 64.5 µl
   - Pfu buffer (10X) 10 µl
   - dNTPs mix (10 mM each) 2 µl
   - DNA template (~40 ng/µl) 2.5 µl
   - Forward primer (10 mM) 10 µl
   - Reverse primer (10 mM) 10 µl

3. Put the reaction mixture in the thermal cycler and start the program.
4. Add 1µl of Pfu polymerase after step 02 (5 minutes at 94°C).
5. Allow the program to proceed.

6. If the PCR product is going to be used for restriction digestion, proceed to purify the DNA by isopropanol precipitation or use the QIAquick PCR purification kit (Qiagen). Resuspend the purified DNA in 30 µl of sterile distilled $H_2O$.

*Colony screening*

- PCR can be used to screen for cells containing a desired DNA fragment without prior plasmid purification. If possible, use a forward primer that anneals with the vector and a reverse primer that anneals with the end of the inserted DNA fragment (or vice versa).

1. Prepare and label 1.5 ml microcentrifuge tubes with 20 µl of sterile LB medium (as many tubes as colonies needed to screen, *n*).
2. Transfer each colony into the corresponding tube with LB medium and resuspend the cells thoroughly.
3. Transfer 10 µl of each cell suspension into an empty labelled PCR tube.
4. Lyse the cells in the PCR tubes by heating for 10 minutes at 95°C.
5. Prepare a bulk PCR reaction mixture for (*n*+2) PCR reactions. Add per reaction:
    - Distilled $H_2O$13 µl
    - PCR buffer (10X) 2 µl
    - dNTPs mix (10 mM each) 0.4 µl
    - Forward primer (10 mM) 2 µl
    - Backward primer (10 mM) 2 µl
    - Taq polymerase (or other DNA polymerase) 0.2µl
6. Aliquot 20 µl of the bulk PCR reaction mixture into *n* PCR tubes and label them.
7. Add 1-2 µl of the lysed cells to the corresponding PCR tube with reaction mixture.
8. Amplify with the standard PCR cycling program described above.
9. Load 5 µl of the PCR reaction into a 2% agarose gel to analyse the PCR products.
10. Use the remaining 10 µl of living cells to inoculate cell cultures.

## Isopropanol precipitation of DNA

1. Add 0.1 volumes of 3 M sodium acetate pH 5.2 to the solution containing the DNA.
2. Add 0.7 volumes of isopropanol.
3. Centrifuge the sample at 13000 rpm for 15 minutes at 4°C.
4. Decant the supernatant.
5. Wash the pellet with 1 ml of 70% ethanol.
6. Centrifuge the sample at 13000 rpm for 10 minutes at 4°C.
7. Decant the supernatant.
8. Air-dry the pellet at 37°C and redissolve the DNA in the desired buffer or in sterile distilled $H_2O$.

## Small scale purification of low-copy plasmid DNA using the QIAprep miniprep kit (Qiagen).

- This is a slightly altered protocol from the one provided by the manufacturer.

1. Select a well-isolated bacterial colony from a freshly-streaked plate to inoculate 10-15 ml of sterile LB medium with the appropriate antibiotic in a sterile disposable 50-ml conical centrifuge tube (Falcon tube).
2. Incubate overnight at 37°C with vigorous shaking.
3. Harvest the cells by centrifugation at 4000-4500 rpm for 15 minutes at 4°C.
4. Decant the medium and invert the tube for 1 minute to remove as much medium as possible.
5. Resuspend the pellet in 500 µl of Tris-EDTA buffer with RNase A (buffer P1) and transfer the resulting suspension to two 1.5 ml microcentrifuge tubes (each with half the amount of the cell suspension).
6. Lyse the resuspended cells by adding 250 µl of the NaOH/SDS solution (buffer P2) to each of the tubes and mix gently.
7. Precipitate proteins, chromosomal DNA and cellular debris by adding 300 µl of the acidic potassium acetate solution (buffer N3) to each of the tubes and mix gently.
8. Centrifuge tubes for 10 minutes at 13000 rpm.

9. Load the supernatants from one of the tubes into a QIAprep spin column by decanting.
10. Centrifuge the spin column for 1 minute at 13000 rpm and decant the flow-through.
11. Load the supernatant from the second tube into the same QIAprep spin column by decanting.
12. Centrifuge the spin column for 1 minute at 13000 rpm and decant the flow-through.
13. Wash remaining enzymes from the column-bound plasmid DNA by loading 0.5 ml of binding buffer (PB) into the column. Centrifuge for 1 minute at 13000 rpm and decant the flow-through.
14. Wash remaining salts from the column-bound plasmid DNA by loading 0.75 ml of ethanol wash buffer (PE) into the column, centrifuge for 1 minute at 13000 rpm and decant the flow-through.
15. Remove the remains of buffer PE by an additional 1 minute centrifugation at 13000 rpm.
16. Place the column in a clean lidless 1.5 microcentrifuge tube (cut the lid with scissors and set aside).
17. Elute the plasmid DNA by loading 50 µl of sterile distilled $H_2O$ to the centre of the column. Let stand for 1 minute and centrifuge for 1 minute at 13000 rpm.

## Restriction endonuclease DNA digestion

- The quantities given here are for digestion of 30 µl of a 50µl-plasmid miniprep (see the small scale purification of low-copy plasmid DNA protocol) or for the digestion of the DNA obtained from a 100 µl-PCR reaction, purified and resuspeded in 30 µl of $H_2O$ (see amplification of DNA fragments protocol).

1. Set up the restriction reactions in ice-chilled 1.5 ml microcentrifuge tubes as follows:
    - $H_2O$ 20 µl
    - Enzyme buffer (10X) 6 µl
    - 30 µl of DNA
    - Restriction enzyme 3 µl
    - BSA (if required by the enzyme) 0.6 µl

Mix well but gently. If setting up a double digest, add 3 µl of the second restriction enzyme and only 17.4 µl of $H_2O$. For restriction enzyme compatibility refer to New England Biolabs catalogue.

2. Incubate for 2 hours at 37°C. If digesting with NdeI add 1.5 µl of enzyme at the beginning and 1.5 µl of enzyme at 20 minutes and do not let the reaction proceed for more that 40 minutes.
3. If the restriction enzymes used generate blunt ends, it is better to remove the 5'-phosphates of the vector to prevent recircularisation. Calf intestinal phosphatase (CIP) and shrimp alkaline phosphatase (SAP) are compatible with most restriction enzyme buffers. Add 1 unit of phosphatase to the vector digest (0.2 units per pmol of DNA) and incubate at 37°C for another 30 minutes.

## Purification of digested DNA by agarose gel electrophoresis

1. Prepare a 1% agarose gel for the purification of plasmid DNA and a 2% agarose gel for the purification of PCR amplification products. Join individual wells with self-adhesive tape in order to make a single large well ~1-1.5 cm wide. Transfer the gel to the electrophoresis apparatus and load the running buffer.
2. Mix the DNA from the digestion reactions with the appropriate amount of sample loading solution and load the large well. Include some uncut vector in a parallel (small) well to see the difference between the linearised and the undigested plasmid.
3. Run the gels at 50 V until the dye is 1 cm from the bottom of the gel.
4. Stain the gel with ethidium bromide.
5. Follow the instructions provided by the manufacturer in the QIAquick gel extraction kit (Qiagen).
6. Elute the DNA from the columns with 30 µl of distilled sterile $H_2O$.

## Ligation of plasmid DNA

1. Estimate the concentration of DNA by running an agarose gel with a DNA mass ladder of the appropriate molecular weight.
2. Mix the appropriate amounts of digested vector and insert in a PCR tube and heat for 5 minutes at 65°C. Cool the tube down on ice and centrifuge for 1 minute at 13 000 rpm.
   - The rule of thumb is to set up ligation reactions with ~0.015-0.03 pmols of vector (50-100 ng of a 5 kbp plasmid) and ~0.2-0.3 pmols of insert (50-75 ng of a 500 bp insert). This is usually around ~10 µl of vector and ~2 µl of insert in a 20 µl reaction. Set up several reactions with 100 ng of vector and different amounts of insert (e.g. 0 pmol, 0.1 pmol, 0.2 pmol, 0.3 pmol).

3. Add the appropriate amount of ligation buffer according to the manufacturer's instructions (usually 2 µl) and distilled $H_2O$ up to 19 µl. If using blunt-ended fragments or DNA digested with NdeI , add PEG to achieve a final concentration of 15%.

   - Make aliquots of the ligase buffer provided with the enzyme. The buffer contains ATP and it is very sensitive to repeated freezing and thawing cycles.

4. Add 1 µl of DNA ligase (3-5 units), mix well by pipetting up and down and incubate overnight at 14-16°C.
5. Transform 50 µl DH5α subcloning efficiency competent cells (Invitrogen) with 10 µl of ligation reaction following the manufacturer's instructions. After transformation and incubation for one hour, harvest the cells by centrifugation at 3000 rpm for 10 minutes. Decant most of the supernatant and plate all the cells in one or two Petri dishes with LB agar and the appropriate antibiotic.

## DNA Sequencing

- Use the Abi Prism dRhodamine (Applied Biosciences) or the Dyenamic ET (Amersham Biosciences) Terminator Cycle Sequencing Kit.

1. Estimate the concentration of DNA by running an agarose gel with a DNA mass ladder of the appropriate molecular weight.
2. Assemble each sequencing reaction in a PCR tube as instructed by the manufacturer. Each reaction consists of:
    - Template DNA
    - Sequencing primer (5 pmol)
    - Sequencing reagents mix (8 μl)
    - Distilled $H_2O$ to a total volume of 20 μl
3. Mix the contents of the tube well by pipetting up and down and centrifuge the tubes briefly.
4. Place the tube in the thermal cycler and run the sequencing program recommended by the manufacturer.
5. Use the DyeEx spin kit (Qiagen) to remove the excess of dye.
6. Air-dry the eluate from the DyEx column or use a centrifugal vacuum concentrator.
7. Take the sample for automated sequencing.
    - Sequencing was done by Dr Athena Nikitopoulou at the Ludwig Institute of Cancer Research using an Abi Prism 377 automated DNA sequencer.

## Tricine denaturing polyacrylamide gel eletrophoresis for proteins

Materials

- 50% (w/v) glycerol
- 40% (w/v) acrylamide / 5% (w/v) bisacrylamide (40%T, 5% C)
- N,N,N'N'-tetramethy-ethylenediamine (TEMED)
- 20% (w/v) ammonium persulphate (APS) freshly prepared
- Gel buffer: 3 M Tris base, 0.3% SDS
- Sample buffer (3X): 50 mM Tris base pH 6.8, 12% glycerol, 4% SDS, 2% 2-mercaptoethanol, 0.01% bromophenol blue
- Cathode buffer (10X): 1 M Tris base, 1 M tricine, 1% SDS. Do not correct pH (should be around 8.25)
- Anode buffer (10X): 2 M Tris base, pH 8.9
- 1% (w/v) SDS
- Coomassie staining solution: 40% methanol, 10% glacial acetic acid, 0.05% (v/w) Coomassie Brilliant blue G-250. Leave stirring overnight and filter.
- Coomassie destaining solution: 40% methanol, 10% glacial acetic acid.

1. Assemble two 0.75 mm-thick gel cassettes.
2. Insert the comb and make a mark on the glass 5 mm below the end of the comb wells.
3. Prepare 12% separating gel by mixing 2.15 g of 50% glycerol, 2.66 ml gel buffer, 2.4 ml of acrylamide/bis (40%T, 5%C), 0.8 ml H2O, 10 µl TEMED and 20 µl of APS. Mix well by inversion.
4. Pour the separating gel mixture into the gel cassettes up to the 5 mm-mark.
5. Overlay the separator gel mixture with 1% SDS and leave to polymerise (15-20 minutes).
6. Decant the 1% SDS from the gel cassettes.
7. Prepare the stacking gel by mixing 0.75 ml of gel buffer, 0.3 ml of acrylamide/bis (40%T, 5%C), 2 ml $H_2O$, 10 µl TEMED and 30 µl of APS. Mix well by inversion.
8. Pour the stacking gel into the gel cassettes and insert the combs and leave to polymerise.
9. Dilute one part of sample buffer with 2 parts of protein sample. Heat the samples at 95°C for 10 minutes.
10. Remove the combs from the gel cassettes and assemble the electrophoresis module.

11. Fill the anode and cathode chambers with the appropriate 1X buffer.
12. Load the samples into the gel wells.
13. Run the gels at 40 mA per gel for around 120 minutes.
14. Remove the gel cassettes from the electrophoresis module and the gels from the cassettes and stain.

## Native polyacrylamide gel electrophoresis for basic proteins

(Adapted from Goldenberg D.P. Analysis of protein conformation by gel electrophoresis in Protein Structure, a practical approach)

Materials

- Acrylamide/ bisacrylamide (30%T, 0.8%C)
- Acrylamide/ bisacrylamide (40%T, 5%C)
- 20% ammonium persulphate freshly prepared
- Separating gel buffer (8X): 12.8 ml acetic acid, 30 ml of $H_2O$, 1 ml TEMED, 1 M KOH to pH 4.0 (around 35 ml), add $H_2O$ to 100 ml
- Stacking gel buffer (8X): 4.3 ml acetic acid, 30 ml of $H_2O$, 0.46 ml TEMED, 1 M KOH to pH 5.0 (around 50 ml), add $H_2O$ to 100 ml
- Electrode buffer (4X): 14.2 g β-alanine, 800 ml of $H_2O$, acetic acid to pH 4.0, add $H_2O$ to 1000 ml
- Tracking solution (5X): 50% (m/v) glycerol, 0.2% (m/v) methyl green

1. Assemble a 1.5mm-thick gel cassette.
2. Prepare the resolving gel solution by mixing in a vacuum flask 4.5ml of $H_2O$, 6ml 30%T, 0.8%C Acrylamide/Bis and 1.5 ml of resolving gel buffer. Put a bung in the flask and connect to vacuum pump for 10 minutes. Add 30 µl of APS 20% and fill the gel cassette to a level 1 cm below the bottom of the comb wells. Overlay with 20% ethanol and leave to polymerise for 60 minutes.
3. Prepare the stacking gel solution by mixing in a vacuum flask 6.5 ml of $H_2O$, 0.5 ml of acrylamide 40% and 1 ml of stacking gel buffer. Put a bung in the flask and connect to vacuum pump for 10 minutes. Add 80 µl of APS 20%. Rinse the top of the separating gel with $H_2O$ and then with stacking gel solution.
4. Fill the gel cassette with the stacking gel solution and insert the comb. Allow to polymerise for 1-2 hrs.

5. Assemble the gel in the electrophoresis chamber, remove the comb and fill the chamber with 1X electrode buffer.
6. Pre-run the gel for 2 hrs at 30 mA per minigel.
7. Prepare the samples and load the gel.
8. Run at 10 mA per minigel for 180 minutes.

## Immuno (Western) blot

- Wet transfer apparatus (Xcell II blot module, Invitrogen)
- Filter paper
- Transfer buffer (25X): 0.3 M Tris base, 2.4 M glycine. Do not adjust pH (Should be ~8.3)
- Transfer buffer (1X): 40 ml of 25X transfer buffer, 200 ml methanol, distilled $H_2O$ to 1 litre.
- TBS (10X): 0.1 M Tris base, 1.5 M NaCl. Adjust pH to 7.5
- TBSTT (5X): 0.1 M Tris base, 2.5 M NaCl, 0.25% Tween-20, 1% Triton X-100
- Blocking solution: 10% skimmed milk power, 0.1% Tween-20 in TBS 1X
- First (mouse pentahistidine) antibody dilution: Use a 1:1000 dilution of antibody solution prepared following the instructions of the manufacturer in TBS 1X with 1% casein or 3% BSA
- Second (goat anti-mouse alkaline phosphatase conjugate) antibody dilution: Use a 1:5000 dilution of antibody solution prepared following the instructions of the manufacturer in blocking buffer
- Alkaline phosphatase (AP) colorimetric detection kit (Novagen)

Transfer of proteins from the SDS-PAGE gel to the membrane

- It is a good idea to run the SDS-PAGE gel with pre-stained markers to monitor the transfer of proteins to the membrane.

1. Cut the protein transfer membrane (PVDF) and two pieces of filter paper to the size of the gel.
2. Immerse the membrane in methanol and leave for 1 minute, then immerse the membrane in transfer buffer and leave until needed.
3. Soak five blotting pads in 500 ml of transfer buffer and squeeze then to remove the air bubbles with a glass pipette.

4. Remove the gel from the glass plates, briefly soak it in transfer buffer and position the transfer membrane over the gel, remove air bubbles.
5. Soak the filter papers in transfer buffer and place one of them on top of the membrane and another on top of the gel, remove air bubbles.
6. Place two soaked blotting pads in the cathode core of the XCell blot module then place the assemble of gel and membrane with the gel facing the cathode (down).
7. Place three soaked blotting pads over the membrane (or enough pads to rise 0.5 cm over the cathode core) and put the anode core at the top.
8. Insert the blot module into the electrophoresis tank.
9. Fill the blot module with just enough transfer buffer to cover the blotting pads.
10. Fill the outer buffer chamber with distilled $H_2O$ to serve as a coolant.
11. Place the lid and connect the unit to the power supply.
12. Run at 25 V for 1-2 hrs (expected current 100 mA), until the pre-stain markers appear on the membrane.

Membrane blocking

1. Place the membrane in a shallow tray.
2. Wash membrane with TBS for 10 minutes, twice.
3. Incubate membrane for 1 hour in blocking solution (Can be left overnight in blocking solution).
4. Wash membrane with TBSTT for 10 minutes, twice.
5. Wash membrane with TBS for 10 minutes.

Antibody binding

1. Incubate membrane in first antibody dilution for 1 hour.
2. Wash membrane with TBSTT for 10 minutes twice.
3. Wash membrane with TBS for 10 minutes.
4. Incubate membrane in second antibody dilution for 1 hour.
5. Wash with TBSTT for 10 minutes 5 times.

Colorimetric detection

1. Prepare developing solution by adding 60 µl of NBT (nitro blue tetrazolium chloride) and 60 µl of BCIP (5-bromo-4-chloro-3-indolyl phosphate) solution to 15 ml of 1X AP buffer.
2. Place the membrane in a clean tray and add the developing solution.
3. Incubate the membrane until colour develops (2-10 minutes).
4. Wash thoroughly with distilled H2O to stop the reaction and let the membrane to dry.
   - Store at room temperature wrapped in plastic.

## Large scale expression of CD5d1 in minimal medium

Protocol is given to prepare one litre of culture.

Growth of initial culture for inoculation.

1. Select a well-isolated bacterial colony from a freshly-streaked plate and inoculate 10 ml of sterile LB medium with 50 µg/ml carbenicillin in a sterile disposable 50-ml conical centrifuge tube (Falcon tube).
2. Incubate for around 8 hrs at 37°C with shaking.
3. Harvest the cells by centrifugation at 3,000 rpm for 10 minutes.
4. Pour the supernatant and resuspend the pellet in some sterile minimal medium.
5. Inoculate the resuspended pellet into 100 ml of minimal medium with 50 µg/ml carbenicillin.
6. Incubate overnight at 37°C with shaking.

Inoculation of culture and induction of expression.

1. Prepare 1 litres of minimal medium in two 2-litre baffled flasks.
2. Harvest the overnight culture by centrifugation at 3,000 rpm for 20 minutes.
3. Decant the supernatant and resuspend the bacterial pellet in 10 ml of sterile minimal medium.
4. Inoculate each one of the 2 flasks with 5 ml of the resuspended cells.
5. Incubate at 37°C at 180 rpm until the absorbance at 600 nm reaches 0.9-1.0 (around 6 hrs).

6. Induce recombinant protein expression with 500 µl of 1 M IPTG (1mM final concentration).
7. Leave expression to proceed for 4 hrs (Expression can be left overnight with a mild decrease in yield).
8. Harvest the cells by centrifugation at 6,000 rpm for 15 minutes.

## Large-scale expression of CD5d1 in LB medium

Protocol is given to prepare one litre of culture.

1. Select a well-isolated bacterial colony from a freshly-streaked plate and resuspend in 1 ml of sterile LB.
2. Use 500 µl of the cell suspension to inoculate each of two 2-litre baffled flasks with 500 ml of sterile LB and 50 µg/ml carbenicillin.
3. Incubate at 37°C and 180 rpm until the absorbance at 600 nm reaches 0.8-1.0 (7-8 hrs).
4. Induce recombinant protein expression with 500 µl of 1 M IPTG (1 mM final concentration).
5. Leave expression to continue overnight.
6. Harvest the cells by centrifugation at 6,000 rpm for 15 minutes.

## Purification of CD5d1 from inclusion bodies

Buffers

- Buffer A: 50 mM $KH_2PO_4$ and 100 mM NaCl. Adjust pH to 7.0 with 5 M NaOH
- Buffer C: 50 mM $KH_2PO_4$, 100 mM NaCl and 2 M urea. Adjust pH to 7.0 with 5 M NaOH
- Buffer E: 6 M GdnHCl or 8 M urea, 0.1 M $NaH_2PO_4$, 0.01 M Tris base. Adjust pH to 8.0 with 5 M NaOH
- Buffer F: 6 M GdnHCl or 8 M urea, 0.1 M $NaH_2PO_4$, 0.01 M Tris base. Adjust pH to 6.3 with 5 M NaOH in using GdnHCl. Buffer pH should be 6.3 without adjusting if using urea.

- Buffer G: 6 M GdnHCl or 8 M urea, 0.1 M $NaH_2PO_4$, 0.01 M Tris base. Adjust pH to 4.5 with 5 M NaOH in using urea. Buffer pH should be 4.5 without adjusting if using GdnHCl.

Preparation of solubilised inclusion bodies.

1. Resuspend the cell pellet in Buffer A. Use 1/50 of the original culture volume.
2. Lyse the cells using the French Pressure Cell three times.
3. Centrifuge the cell suspension for 30 minutes at 20,000 rpm.
4. Decant the supernatant and resuspend the pellet in Buffer C using the homogenizer. Use 1/100 of the original culture volume.
5. Centrifuge for 40 minutes at 20,000 rpm.
6. Decant the supernatant and resuspend the pellet in buffer E using the homogenizer. Use 1/100 of the original culture volume.
7. Incubate for two hours at RT (Gdn HCl) or 30°C (urea).
8. Centrifuge for 60 minutes at 20,000 rpm.
9. Decant the supernatant and adjust the pH to 8.0 if necessary.
10. Repeat twice if the sample is isotope-labelled.

Gravity-driven immobilised metal ion affinity chromatography.

1. Use 5 ml of Ni-IDA agarose resin (His Bind resin, Novagen) in a 30-ml or bigger gravity-driven chromatographic column.
2. Equilibrate with 5 CV of Buffer E.
3. Load 3 CV of solubilised inclusion bodies.
4. Wash with 5-10 CV of Buffer F or until the Bradford assay (RioRad) is negative for 50µl of flow-through.
5. Elute with 5 CV of Buffer E or until the Bradford assay in negative for 50µl of flow-through.
6. Regenerate the column by washing with 3 CV of strip buffer, 5 CV of $H_2O$, 5 CV 50 mM of $NiSO_4$ and 10 CV of $H_2O$.

Concentration and reduction.

1. Concentrate the eluate from the metal ion affinity chromatography by centrifugal ultrafiltration (Vivaspin 20) until the protein concentration is around 20 mg/ml as measured with the Bradford assay (BioRad).
2. Set the pH to 8.0 with diluted NaOH and add enough solid DTT to achieve a final concentration of 100 mM.
3. Leave overnight at RT with gentle mixing.

Size-exclusion chromatography.

1. Run a SEC with the following specifications:

   Matrix: Superdex 75 (Amersham Biosciences)
   Column: XK16
   Mobile phase: Buffer G
   Flow rate: 0.5 ml/min
   Injection: ~ 20 mg of protein or less than 2 ml
   Fractions: 2.5 ml fractions between 0.3 and 0.60 CV

   - Proteins start to elute around 0.32 CV. The peak corresponding to monomeric CD5d1 should elute between 0.40 and 0.55 CV. Low molecular weight components elute between 0.89 and 1.2 CV.

2. Pool the fractions of monomeric CD5d1 and concentrate to more than 10 mg/ml using centrifugal ultrafiltration (Vivaspin). The samples at this point are ready for *in vitro* folding.

   - Samples should never be stored in 8 M urea. If it is not possible to fold immediately, the buffer has to be changed to GdnHCl or 0.1% TFA for storage.

## In vitro oxidative folding of CD5d1

- Folding buffer: 0.1 M Tris base, 0.6 M arginine, 20% glycerol, pH 9.0 Filter and degas by stirring the buffer while connected to the water-driven vacuum. Equilibrate at 4°C.
- Cysteine (100X): 0.5 M cysteine
- Cystine (100X): 0.05 M cystine. In order to dissolve add a few drops of 5 M NaOH

- Buffer B: 50 mM $KH_2PO_4$ and 200 mM NaCl, 1 mM EDTA, 0.1% Sodium azide. Adjust pH to 5.0 with 5 M NaOH

1. Add the necessary amount of cysteine and cystine to the folding buffer.
2. Measure protein concentration of the unfolded monomeric sample by absorbance at 280 nm. The theoretical extinction coefficient for reduced CD5d1 is 27880 $M^{-1}cm^{-1}$ and the molecular weight is 14721 g/mol. The concentration has to be above 10 mg/ml.
3. While gently mixing the in vitro folding buffer with a stirring magnet, add enough protein to achieve a final concentration of 0.5 mg/ml. If time allows, add smaller aliquots every hour.
4. Incubate at 4°C for 24 hrs with gentle stirring.
5. Remove the precipitate by centrifugation at 4500 rpm for 15 minutes followed by syringe-driven filtration.
6. Concentrate the folded protein solution to ~30 ml using a stirred ultrafiltration cell (Amicon).
7. Dialyse the concentrate against Buffer B for 24 hours with 3-4 buffer changes.
8. Concentrate using a stirred ultrafiltration cell and/or centrifugal ultrafiltration units.
9. Purify by size-exclusion chromatography with the following specifications:

   Matrix: Superdex 75 (Amersham Biosciences)
   Column: XK16
   Mobile phase: Buffer B
   Flow rate: 0.5 ml/min
   Injection: 2 ml of in vitro folding concentrate
   Fractions: 2.5 ml fractions

10. Concentrate using centrifugal ultrafiltration units.

## In vitro folding screen for mutants

- Enough protein to screen a mutation is obtained by growing 500 ml of cells in LB.
- Proceed to solubilise and purify the inclusion bodies as explained for CD5d1.

- Load all the GdnHCl solubilised protein (should be 10-15 ml) into 5ml of Ni-IDA resin and elute as described for the purification of CD5d1.

1. Add diluted NaOH to the eluate to set the pH to 8.5. Concentrate to less than 2 ml by centrifugal ultrafiltration (Vivaspin 20). The final concentration should be above 10 mg/ml, as measured with the Bradford assay.
2. Add solid DTT to achieve a final concentration of 0.1 M and incubate with gentle stirring overnight at room temperature.
3. Size-exclusion chromatography.
   - Specifications

     Column: Superdex 75. XK16

     Mobile phase: Buffer G

     Flow rate: 0.5 ml/min 0 to 0.6 CV; 1.0 ml/min from 0.6 to 1.2 CV

     Injection: ~ 20 mg of protein or less than 2 ml

     Fractions: 2.5 ml fractions between 0.3 and 0.60 CV

   Pool fractions between 0.4 and 0.45 CV.
   - Using the 55cm XK16 column this corresponds to 45 to 50 ml (fractions A5 and A6) Concentrate until the protein concentration is above 10 mg/ml. This is usually a 10x or more concentration. Measure protein concentration using the Bradford assay.
4. Prepare six universal tubes with 5-10 ml of a degassed solution of 0.6 M arginine, 20% glycerol and 0.1 M Tris at pH 9.0. The chosen volume depends on the amount of purified protein obtained. A 5ml screen requires 5 mg of protein. Equilibrate at 4°C.
5. Prepare 5 ml of stock solutions of cysteine and cystine and add to the folding buffer.
6. Add the reduced protein to the folding buffer to achieve a final concentration of 0.1mg/ml in three of them and 0.2 mg/ml in the other three. For the 0.2 mg/ml samples, add the protein in two steps, allowing some time in between the additions.
7. Incubate at 4°C for 24 hrs with gentle stirring.
8. Dialyse all the samples overnight against Buffer B.
9. Run RP-HPLC
   - Specifications

     Column: C4 100 mm

     Mobile phase A: 5% acetonitrile, 0.1% TFA

Mobile phase B: 95% acetonitrile, 0.1% TFA

Flow rate: 2 ml/min

Injection: 2 ml of in vitro folded protein

Gradient: 0-45% B in 15 minutes

## Folding screen for model proteins

1. Denaturation and reduction of disulphide bonds.

Prepare 3 ml of a 10 mg/ml protein solution in buffer E using GdnHCl.

Determine protein concentration by measuring absorbance at 280 nm.

Take a 60 µl aliquot and store at 4°C.

Add solid DTT to the rest of the protein solution to achieve a final concentration of 100 mM DTT.

Incubate overnight at 4°C with gentle mixing.

Take a 60 µl aliquot and store at 4°C.

Extinction coefficients of oxidised species at 280 nm:

Lysozyme 38460 $M^{-1}cm^{-1}$. Molecular weight 14.3 kg/mol

Ribonuclease A 8160 $M^{-1}cm^{-1}$. Molecular weight 13.7 kg/mol

2. Refolding.

Prepare the *in vitro* folding buffers as indicated in the table.

Pour 5 ml of each buffer in a separate 30ml universal tube. Add 100 µl of 0.5 M cysteine and 100 µl of 0.05 M cystine. Place the 4°C buffers in the fridge and wait for them to equilibrate.

Dilute the protein solution in each of the *in vitro* folding buffers at the appropriate concentration.

Incubate at the indicated temperature with gentle mixing for 24-36 hrs.

3. Dialysis.

Centrifuge the tubes at 4000 rpm for 10 minutes.

Dialyse 4 ml of each run against 100 volumes of Buffer B at 4°C overnight.

Use ~15 cm of narrow dialysis tubing (23 mm width, Spectra/Por 1) for each one.

4. RP-HPLC analysis.

Record the final volume of each dialysed sample.

Dilute 30 µl of the oxidised sample into 3 ml of buffer B.

Dilute 30 µl of the reduced sample into 3 ml of buffer E or 5%AcCN/0.1% TFA.

Run the ten samples using a 1 ml loop and injecting 2.5-3 ml.

If the amount of soluble protein in the buffers with initial protein concentration of 0.5 mg/ml is too high, dilute 1:5 with buffer B.

## Materials

*Amersham Biosciences (GE Healthcare)*
Hybond P (PVDF) protein transfer membrane
Superdex 75 prep grade SEC medium

*BDH AnalaR*
NaCl
$NaH_2PO_4$
NaOH
$KH_2PO_4$
Urea
D(+)-Glucose
Calcium chloride (1 M) solution
Sodium dodecyl sulphate
2-mercaptoethanol

*BDH Electran*
Acrylogel 5 solution

*Biorad*
Protein assay (Bradford)

*Invitrogen*
Mark-12 Molecular weight standard
Subcloning efficiency DH5α competent cells
Low and high DNA mass ladders.

*Melford Labs Ltd.*
Dithiothreitol (DTT)
Isopropyl-β-D-thiogalactopyranoside
Carbenicillin, disodium salt
Kanamycin, monosulphate
Guanidine hydrochloride 99%

*New England Biolabs*
Endonucleases
T4 DNA ligase
Calf intestine phosphatase
Taq polymerase
1kb DNA ladder
100 bp DNA ladder

*Novagen*
HisBind resin
Goat anti-mouse IgG AP conjugate
AP detection reagent kit

*Promega*
Pfu polymerase

*Qiagen*
QIAquick Gel Extraction kit
QIAquick PCR purification kit
QIAprep Spin Miniprep Kit
DyeEx 2.0 Spin kit
HisTag Monoclonal Antibody

*Sigma/Aldrich/Fluka*

Acetonitrile (99.93%) HPLC grade (Aldrich)
Ammonium persulphate
Bovine serum albumin (98%)
Cysteine
Cystine (Fluka)
Glycerol
Guanidine hydrochloride (98 and 99%)
L-Arginine Hydrochloride (98%)
N,N,N'N'-tetramethy-ethylenediamine
Nickel Sulfate · 6 $H_2O$
Tricine
Trifluoroacetic acid HPLC grade (Fluka)
Trizma-Base

*Spectrum*

Spectra/Por 1 membrane
MWCO 6-8000
Spectra/Por 3 membrane
MWCO 3000

*Stratagene*

QuickChange site-directed mutagenesis kit

*Vivascience*

Vivaspin concentrators (6 and 20 ml)
3000 and 5000 MWCO

## Equipment

*American Instrument Company*
French Pressure Cell Press

*Amersham Biosciences (GE Healthcare)*
XK16/70 empty chromatography column
Äkta purifier chromatography system

*BioRad*
Disposable chromatography columns
Mini Protean II electrophoresis system
Mini Protean III electrophoresis system
PowerPac 3000 power supply
Bio-Silect SEC 125-5 Column

*Eppendorf*
Centrifuge 5810R

*Heraeus*
Labofuge 400R

*Invitrogen*
Xcell II blot module
Xcell SureLock mini-cell

*Millipore*
Amicon stirred cell for ultrafiltration
Amicon YM ultrafiltration discs 3 kDa

*MWG-Biotech*
Primus thermocycler

*Perceptive BioSystems*
Biocad chromatography system

*Sanyo*
Ultrasonic disintegrator Soniprep 150 with exponential probe 10microns amplitude

*Sorvall*
Superspeed centrifuge 5RB
SLA-3000 rotor
SS-34 rotor

# Appendix B

## Oligonucleotide sequences

### CD5d3 expression vector

*CD5D3_FOR_NdeI*

5'-GAT GAT GAT GAT CAT ATG GGT TTC CAG CCC AAG GTG-3'

*CD5D3_REV_stop_XhoI*

5'-GAT GAT GAT GAT CTC GAG CTA GCC TGC GGG GTT TGG ATC-3'

### Site-directed mutagenesis

*L26K*

5'-CCG CGC GGC AGC CAT ATG CGG AAA AGC TGG TAT GAC CCA GAT TTC-3'
5'-GAA ATC TGG GTC ATA CCA GCT TTT CCG CAT ATG GCT GCC GCG CGG-3'

*W28R*

5'-GGC AGC CAT ATG CGG CTC AGC CGT TAT GAC CCA GAT TTC CAG GC-3'
5'-GC CTG GAA ATC TGG GTC ATA ACG GCT GAG CCG CAT ATG GCT GCC-3'

*V88N*

5'-GC CAG CGG CTG AAC TGT GGG AAC CCC TTA AGC CTT GGC CCC TTC-3'
5'-GAA GGG GCC AAG GCT TAA GGG TGG CCC ACA GTT CAG CCG CTG GC-3'

*V88D*

5'-GC CAG CGG CTG AAC TGT GGG GAT CCC TTA AGC CTT GGC CCC TTC-3'
5'-GAA GGG GCC AAG GCT TAA GGG ATC CCC ACA GTT CAG CCG CTG GC-3'

*V97K*

5'-GCC TTG GCC CCT TCC TTA AAA CCT ACA CAC CTC AGA GCT CAA TC-3'
5'-GAT TGA GCT CTG AGG TGT GTA GGT TTT AAG GAA GGG GCC AAG GC-3'

*M124Q*

5'-GC AGC CAC AGC AGA AAT GAC CAA TGT CAC TCT CTG GGC CTG ACC-3'
5'-GGT CAG GCC CAG AGA GTG ACA TTG GTC ATT TCT GCT GTG GCT GC-3'

*M124N*

5'-GC AGC CAC AGC AGA AAT GAC AAC TGT CAC TCT CTG GGC CTG ACC-3'
5'-GGT CAG GCC CAG AGA GTG ACA GTT GTC ATT TCT GCT GTG GCT GC-3'

*G129S*

5'-GAC ATG TGT CAC TCT CTG TCC CTG ACC TGC TTA GAA TAG GG-3'
5'-CC CTA TTC TAA GCA GGT CAG GGA CAG AGA GTG ACA CAT GTC-3'

# Appendix C

## Experimental details of NMR experiments

### $^{13}C/^{15}N$ V88D/V97K *h*CD5d1 NMR experiments at 25°C

| | $^{1}H$ | | $^{13}C$ | | $^{15}N$ | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 1024 (102ms) | 5000 | | | 256 | 1700 (150ms) | 16 |
| HNCA | 1024 (102ms) | 5000 | 64 (16ms) | 4050 | 32 | 1700 (19ms) | 16 |
| HNCACB | 1024 (102ms) | 5000 | 64 (6ms) | 10200 | 32 | 1700 (19ms) | 24 |
| CBCA(CO)NH | 1024 (102ms) | 5000 | 70 (7ms) | 10200 | 34 | 1700 (19ms) | 16 |
| H(CCO)NH-TOCSY (12ms mixing time) | 1024 (98ms) | 5200 | 64 (17ms) | 3800 | 32 | 1700 (19ms) | 8 |

Experiments run at 600 MHz except HCCONH run at 500MHz

## $^{13}C/^{15}N$ V88D/V97K *h*CD5d1 NMR experiments at 35°C

| | $^{1}H$ | | $^{13}C/^{1}H$ | | $^{15}N$ | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 1024 (102ms) | 5000 | | | 256 (150ms) | 1700 | 16 |
| HNCA | 1024 (120ms) | 4200 | 70 (17ms) | 4050 | 32 (19ms) | 1700) | 16 |
| HNCACB | 1024 (102ms) | 5000 | 64 (8ms) | 10200 | 32 | 1700 | 32 |
| (H)C(CO)NH-TOCSY (12ms) | 1024 (102ms) | 5000 | 80 | 10200 | 32 | 1700 | 16 |
| HA(CA)NH | 1024 | 5000 | 40 | 1700 | 32 | 1700 | 16 |
| HA(CACO)NH | 1024 | 5000 | 40 | 1700 | 32 | 1700 | 16 |

Experiments run at 600 MHz

## $^{13}C/^{15}N$ V88D/V97K *h*CD5d1 NMR experiments at 45°C pH 5

| | $^{1}H$ | | $^{13}C$ | | $^{15}N$ | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 1024 | 5000 | | | 256 | 1700 | 8 |
| HNCO | 1024 | 4200 | 64 | 1600 | 32 | 1700 | 8 |
| HNCA | 1024 | 4200 | 66 | 4050 | 32 | 1700 | 12 |
| HNCACB | 1024 | 5000 | 57 | 8200 | 32 | 1500 | 24 |
| CBCA(CO)NH | 1024 | 5000 | 57 | 8200 | 32 | 1500 | 26 |
| HA(CACO)NH | 1024 | 4200 | 64 | 1800 | 32 | 1700 | 8 |

Experiments run at 600 MHz except CBCACONH/HNCACB run at 500MHz

## $^{13}C/^{15}N$ V88D/V97K *h*CD5d1 NMR experiments in $D_2O$ at 45°C pH 5

| | $^1H$ | | $^{13}C$ | | $^1H$ | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 1024 | 5500 | 150 | 9900 | | | 16 |
| CT-HSQC | 1024 | 5500 | 256 | 9900 | | | 48 |
| HCCHTOCSY(20ms) | 1024 | 5500 | 96 | 10000 | 64 | 3800 | 8 |
| NOESY (100ms) | 1560 | 7800 | 58 | 6040 | 100 | 7800 | 8 |

Experiments run at 600 MHz except NOESY run at 800 MHz

## $^{13}C/^{15}N$ V88D/V97K *h*CD5d1 NMR aromatic experiments in $D_2O$ at 45°C and pH 5

| | $^1H$ | | $^{13}C$ | | $^1H$ | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 2048 | 8000 | 40 | 4200 | | | 16 |
| CT-HSQC | 2048 | 8000 | 128 | 4200 | | | 32 |
| HCCHTOCSY(12ms) | 2048 | 8000 | 32 | 4200 | 48 | 2400 | 8 |
| NOESY | 512 | 3200 | 32 | 4200 | 86 | 6400 | 16 |

## $^{15}$N V88D/V97K *h*CD5d1 NMR experiments at 45°C and pH 5

| | $^{1}$H | | $^{15}$N | | $^{1}$H | | |
|---|---|---|---|---|---|---|---|
| Experiment | Points | Sweep Width | Points | Sweep Width | Points | Sweep Width | Transients |
| HSQC | 1024 | 5000 | 128 | 1250 | | | 4 |
| TOCSY | 1024 | 5000 | 32 | 1250 | 100 | 5100 | 8 |
| NOESY | 1024 | 5000 | 32 | 1500 | 128 | 6800 | 16 |
| HNHA | 1600 | 6400 | 36 | 1250 | 128 | 3600 | 8 |
| HNHB-2D | 1024 | 6400 | 64 | 1250 | | | 16 |
| HNHB-3D | 1024 | 6400 | 32 | 1250 | 64 | 5200 | 16 |
| T1 | 1024 | 6400 | 100 | 1250 | | | 32 |
| T2 | 1024 | 6400 | 100 | 1250 | | | 32 |
| NOE | 1024 | 6400 | 100 | 1250 | | | 80 |

Relaxation times:
**T1** 10, 10, 151, 302, 453, 603, 603, 905, 905 ms; **T2** 16, 16, 80, 143, 207, 239, 239 ms

NOE: 0, 3sec

Experiments run at 500 MHz except NOESY-HSQC run at 600MHz